EyeGuard: A Computer Vision-Based Approach to Detecting Shoplifting in Retail Stores

Author	Afifa Masood, Rabail Waseem, Hamza Arshad, Urooj Fatima
Published Date	February 17, 2025
Instructor	Dr. Muhammad Farooq

Abstract

Shoplifting is a major concern for retail businesses, leading to financial losses and security challenges. Traditional surveillance systems require continuous human monitoring, which is impractical for real-time detection due to the vast amount of video data. This research proposes an intelligent shoplifting detection system that leverages a hybrid deep learning approach. Along with Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) for temporal sequence analysis, the model was also trained at the frame level, significantly improving detection accuracy. Additionally, both supervised and unsupervised learning techniques were utilized to enhance model robustness. Unlike existing approaches, our dataset was created using real-world surveillance footage from stores in Islamabad and Lahore, capturing both crowded and simple background scenarios to ensure adaptability across different environments. The results demonstrate superior performance, making the proposed system a promising solution for automated retail security. Future work will focus on real-time deployment and optimizing the model for resource-constrained environments.

Keywords: Shoplifting; video surveillance; classification; features; convolutional neural network; RNN

1. Introduction

As a common problem faced by stores, shoplifting affects not just businesses but also the community, making it a big concern for shop owners, police, and lawmakers. It not only leads to financial losses for store owners but also creates challenges for security agencies and the judicial system. Shoplifting typically involves concealing items in pockets, bags, or clothing and

leaving the store without paying. According to the National Association for Shoplifting Prevention, approximately 1 in 11 people engage in shoplifting, and offenders are caught only once in every 48 thefts.[1]

There are numerous methods of shoplifting, most of which revolve around concealing merchandise and executing an abnormal sequence of actions. While a human observer can retrospectively identify suspicious behavior in recorded footage, real-time monitoring becomes increasingly challenging as the number of shoppers—and potential shoplifters—grows. Modern retail spaces, shopping malls, and public places are now equipped with extensive surveillance networks to enhance security. However, the sheer volume of video data generated by these systems makes manual monitoring impractical for security personnel.[2]

To address this challenge, artificial intelligence has emerged as a powerful tool for automating video surveillance and detecting shoplifting incidents with greater accuracy. Deep learning techniques, particularly those focused on anomaly detection and action classification, have proven highly effective in identifying suspicious activities in real-time.[1,4-7] This research introduces an AI-driven shoplifting detection system that combines supervised and unsupervised learning approaches. The model is trained on real-world surveillance footage collected from stores in Islamabad and Lahore, ensuring adaptability across various retail environments.

2. Theoretical Frameworks

The most common applications of artificial intelligence in video monitoring include motion detection, face recognition, inactivity detection, and anomaly behavior detection. Several studies have explored various AI-based techniques to detect shoplifting. In [9], researchers developed a classification model using the Jubatus plug-in to extract image features and assess anomalous customer behavior. Their system classified surveillance video data using a linear classifier and kNN classifier to estimate the probability of shoplifting. Tsushita and Zin [10] proposed an algorithm that divides video frames into eight sections and monitors speed changes to detect theft and violence. Another approach [11] used genetic algorithms to generate neural networks for classifying human behavior based on posture changes in video sequences. Similarly, [12] implemented a deep learning model that processes extracted video features using long-term convolutional memory layers for theft detection. In [13], MobileNet deep learning models were used for feature extraction, combined with an improved particle swarm optimization algorithm to classify actions in video sequences.

Convolutional neural networks (CNNs) have gained popularity for theft detection, especially 3D CNNs, which extract both spatial and temporal video features. Several studies have employed 3D CNNs for object recognition, human action recognition, and gesture detection. A pre-trained C3D model [14] was used for anomaly detection, focusing on theft, fights, and traffic incidents. Similarly, [16] classified video frames into normal and shoplifting categories using a 3D CNN network and a fully connected neural network. Another study [17] applied the frame difference method to detect suspicious actions using a 3D convolutional model. In [18], researchers

analyzed violent events in surveillance videos using a deep learning model without relying on manual feature extraction. Studies like [19] demonstrated that combining CNNs with recurrent neural networks (RNNs) improved classification performance. The use of 3D CNNs combined with LSTM networks [20, 21] further enhanced human action recognition by considering pose, illumination, and environment variations. Lastly, an expert system [22] used CNNs and long-term memory modules to detect shoplifting based on motion and appearance analysis in video sequences.

Frame-level feature extraction was utilized to train the human activity recognition model, where individual video frames were processed independently before being aggregated for final classification. This approach ensures that spatial features from each frame are effectively captured using a Convolutional Neural Network (CNN). CNNs play a crucial role in this technique as they enable hierarchical feature learning, allowing the model to detect edges, textures, and complex structures essential for recognizing human actions [25]. The training process involved extracting features from each frame using a CNN architecture. These extracted features were then used to analyze temporal dependencies, often through sequential models like Recurrent Neural Networks (RNNs) or Gated Recurrent Units (GRUs) [26]. CNN was specifically chosen for its strong spatial feature extraction capability, which improves the robustness of the action recognition model. This method eliminates the need for handcrafted features and makes the system more scalable for large datasets [27]. The combination of frame-level processing with CNNs ensures that each frame contributes to the overall recognition of an action while maintaining computational efficiency [28].

A hybrid approach combining YOLO (You Only Look Once) and ResNet was employed to enhance the detection of shoplifters through real-time object detection and deep feature extraction. YOLO was utilized for its capability to generate precise bounding boxes around potential shoplifters in surveillance footage, enabling rapid localization of suspicious activities [29]. Meanwhile, ResNet34 was integrated for its robust feature extraction, allowing the model to analyze spatial patterns and distinguish between normal and suspicious behaviors with higher accuracy [25]. This combination ensures an efficient balance between speed and accuracy, making it suitable for real-time security applications. Currently, the model successfully detects shoplifting behavior to a certain extent, demonstrating promising results. However, future improvements will focus on refining detection accuracy by incorporating advanced training techniques, increasing dataset diversity, and optimizing hyperparameters. Further enhancements will also include fine-tuning the YOLO framework to reduce false positives and integrating sequential models to capture temporal dependencies in shoplifting behavior. With continuous improvements, this system is expected to provide a more reliable and intelligent surveillance solution for retail security [26][27].

3. Problem Statement

Summarizing the research review, we proposed shoplifting detection system focuses on *frame-level analysis*, where each video frame is individually processed to identify suspicious behavior. A *Convolutional Neural Network (CNN)* plays a crucial role in extracting spatial

features from each frame, capturing details such as object movement, posture, and hand positioning. This allows the model to detect potential shoplifting actions even before analyzing temporal patterns. Each frame is assigned a label—*"shoplifting"* or *"normal"*— creating a structured dataset that enhances the model's ability to learn fine-grained behavioral distinctions. By emphasizing frame-level classification, the system ensures real-time detection and immediate response, making it highly effective for surveillance applications. While a Recurrent Neural Network (RNN) is incorporated to understand sequential dependencies, the primary strength of this model lies in its ability to independently classify frames using CNN-based feature extraction. This approach improves detection accuracy and ensures that even a single frame can provide critical insights for preventing theft incidents.

4. Methodology

4.1. *Dataset Preparation*

A **custom dataset** was created by collecting real surveillance footage from retail stores in **Islamabad and Lahore**. Our dataset was tailored to ensure relevance and diversity in customer actions. A total of **300 videos** were recorded, consisting of **balanced instances of both shoplifting and non-shoplifting scenarios**. Each video was captured at different time intervals to include variations in lighting, crowd density, and store layouts.

To enhance the dataset, each video was **segmented into shorter clips**, ensuring the model could learn fine-grained behavioral patterns. The final dataset was categorized into two distinct classes: **Class 0 (normal) and Class 1 (shoplifting), with an equal number of samples in each class**. This balanced dataset provided a strong foundation for training the model to distinguish between normal shopping activities and potential shoplifting incidents with higher accuracy.

Figure 1(a) shows a frame from a video where shoppers are simply selecting products, while Figure 1(b) depicts a frame from a video capturing an instance of shoplifting.

(a)

(b)

Figure 1. Video frames: (**a**) normal (**b**) shoplifting

4.2. *Selection of Neural Network for Classification*

The two most widely used deep learning architectures for video classifications are convolutional and recurrent neural networks. CNNs are mainly used to learn spatial information from a video, whereas RNNs are used to learn temporal information. These network architectures are typically

used for different purposes. However, since video data contains both spatial and temporal components, a combined approach using both architectures is often necessary.[3]

Convolutional neural networks (CNNs) are a type of neural network designed for image recognition and object detection through computer vision. CNNs operate by extracting essential features from images while reducing dimensionality but preserving critical information. This is achieved through multiple layers, including convolutional layers, pooling layers, fully connected layers, and normalization layers. The convolutional layers simulate a neuron's response to a visual stimulus by applying a convolution operation to the input, generating feature maps that are passed to subsequent layers.[3]

Recurrent neural networks (RNNs), on the other hand, are specialized for processing sequential data. Unlike CNNs, which are designed for spatial scaling, RNNs are adept at handling variable-length sequences, making them ideal for capturing temporal dependencies in video frames. A recurrent neuron's output at a given time step depends on its previous inputs, essentially giving it a memory-like capability. Among RNN variants, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are widely used for mitigating long-term dependency issues. GRUs, introduced in 2014, are computationally more efficient and require fewer parameters compared to LSTMs, making them faster to train.[5]

Recent advancements in transformer-based architectures, such as Vision Transformers (ViTs) and Video Swin Transformers, have further improved video classification performance by leveraging self-attention mechanisms to capture long-range dependencies in video data [12]. These models have shown state-of-the-art results on various benchmark datasets [13].

In this study, we employed a hybrid CNN-RNN approach for video classification, focusing on both frame-level and video-level feature extraction. Specifically, we used a convolutional neural network (CNN) for spatial feature extraction and an RNN to process the extracted feature sequences. Unlike GRU-based architectures, our approach utilized a standard RNN structure to retain simplicity and efficiency.

For spatial feature extraction, we leveraged CNN models pretrained on the ImageNet dataset. Two specific models were used: MobileNetV3Large and ResNet34. MobileNet was selected due to its optimized architecture for low-resource environments, ensuring real-time performance. It is particularly suited for deployment in applications requiring fast inference times, such as surveillance systems. ResNet34, a deeper architecture, was also employed to enhance feature extraction, leveraging its residual connections to improve learning efficiency.

Feature extraction was performed at the frame level using CNNs, where each video frame was processed independently to obtain a feature vector representation. These extracted features were then fed into the RNN model for temporal sequence analysis at the video level. The combined CNN-RNN architecture enabled robust video classification by capturing both spatial and temporal dependencies effectively.

Transfer learning played a crucial role in this model. By utilizing pretrained CNNs, we accelerated the training process while benefiting from high-quality feature representations.[27] The models were loaded with pretrained weights from the ImageNet dataset, ensuring optimal performance without the need for extensive retraining [28]. The CNNs were used without their fully connected layers, allowing the extracted feature maps to serve as inputs for the subsequent RNN module [29].

In summary, our approach integrates CNN-based spatial feature extraction at the frame level and RNN-based temporal sequence processing at the video level. Unlike traditional methods that rely on GRUs or LSTMs, our model maintains a straightforward CNN-RNN architecture to balance accuracy and computational efficiency [30]. The use of MobileNetV3Large and ResNet34 further enhances the system's robustness, making it suitable for real-time applications such as shoplifting detection in surveillance video streams [31].

4.3. *Evaluation of Classification Performance*

To evaluate the obtained results during a binary classification, an error matrix with true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values was used. The confusion matrix provides a comprehensive summary of the classification performance, making it a valuable tool in shoplifting detection.

Classification Metrics

Accuracy

Accuracy represents the fraction of correct predictions made by the classifier and is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

However, it is important to note that accuracy is not always an informative metric, especially when dealing with imbalanced datasets where one class significantly outnumbers the other.

Precision

Precision quantifies the proportion of correctly predicted positive instances among all instances classified as positive:

$$Precision = \frac{TP}{TP+FP}$$

Precision is crucial in shoplifting detection, as it helps minimize false positives, ensuring that non-shoplifters are not wrongly classified as shoplifters.

Recall (Sensitivity)

Recall measures the classifier's ability to identify actual positive instances and is calculated as:

$$Recall = \frac{TP}{TP+FN}$$

A high recall value indicates that the model is effective in detecting shoplifting cases, even if it sometimes misclassifies non-shoplifting instances.

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This metric is particularly useful when working with imbalanced datasets, as it ensures a balance between precision and recall.

Confusion Matrix

The confusion matrix is a tabular representation that summarizes the classification outcomes:

Actual / Predicted	Shoplifting (Positive)	Normal (Negative)
Shoplifting (Positive)	TP	FN
Normal (Negative)	FP	TN

This matrix allows us to analyze the types of errors the model makes and adjust it accordingly to improve performance.

5. *Experiments*

We developed an advanced algorithm to tackle the shoplifting recognition problem as a classification challenge. The methodology involved multiple stages, including data collection, preprocessing, feature extraction, and model training. Each stage was crucial in ensuring the model’s accuracy and robustness in detecting shoplifting behavior from surveillance videos.

Data Collection and Augmentation

The initial phase involved gathering a dataset of surveillance footage from retail stores. Since the dataset was imbalanced, with a significant majority of instances labeled as normal customer behavior, we applied undersampling to achieve balance. Specifically, we retained 155 instances from each class, resulting in an initial dataset of 310 video fragments. However, given the complexity of human action recognition, this dataset size was insufficient for robust model training. To address this limitation, we employed data augmentation techniques. Each video fragment was horizontally mirrored to double the dataset size to 620 instances. Further, we introduced slight rotational variations by rotating frames 5 degrees left and right, ultimately expanding the dataset to 1860 video fragments. This augmentation process helped improve the model's generalization ability by providing diverse variations of shoplifting behaviors.

Data Preprocessing and Labeling

Following data augmentation, we proceeded with preprocessing to ensure consistency and quality in the dataset. Each video fragment was labeled into two distinct categories: shoplifting incidents (Class 1) and normal customer behavior (Class 0). Subsequently, the videos were resized to a standardized resolution of 224×224 pixels to match the input requirements of the convolutional neural network used in feature extraction. To further streamline the data, each video fragment was divided into frames at a rate of 10 frames per second over a duration of 3 seconds, yielding sequences of 30 frames per video fragment. The dataset was then split into two subsets: 1302 fragments for training and 558 for testing, ensuring an optimal balance between training efficiency and evaluation reliability.

Feature Extraction Using Deep Learning CNN-RNN

Feature extraction played a critical role in enhancing the model's ability to identify relevant patterns. To achieve this, we utilized the MobileNetV3Large convolutional neural network, a lightweight yet highly effective deep learning model pre-trained on the ImageNet-1k dataset. MobileNet was chosen due to its efficiency in extracting complex spatial features from images while maintaining computational efficiency. The extracted feature maps from each frame served as input to the subsequent recurrent neural network (RNN) model, facilitating the recognition of sequential dependencies in human actions.

To further refine our approach, we combined CNNs and RNNs, leveraging MobileNet for spatial feature extraction and recurrent network for temporal modeling. This hybrid approach ensured that both spatial and temporal dependencies were effectively captured, allowing the model to detect suspicious behavior patterns with higher accuracy. The CNN processed each frame independently, extracting discriminative visual features, while the RNN network learned the sequential relationships between frames, making the system adept at recognizing complex human movements over time.

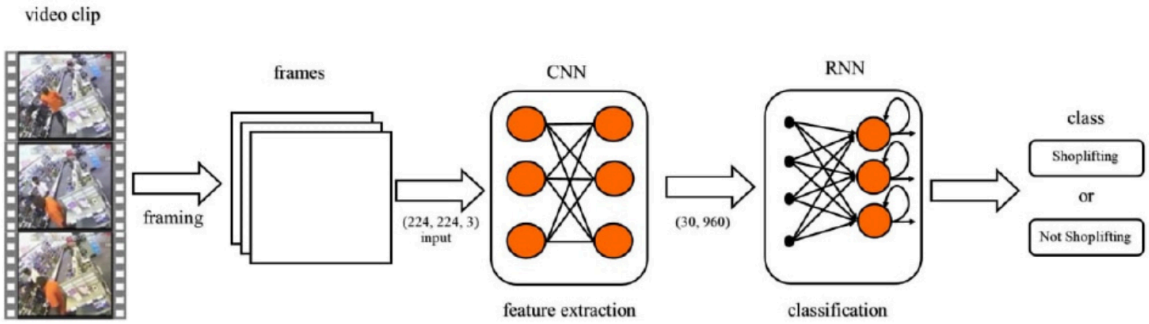

Figure 2. The main stages of the classification algorithm CNN-RNN

Another Experiment on Frame-Level Feature Extraction

An essential aspect of our approach was frame-level feature extraction, which allowed the model to analyze fine-grained spatial details within each video sequence. Instead of treating entire video fragments as single units, we processed them at the frame level, ensuring that each frame contributed to the overall understanding of the action. Each frame was passed through MobileNetV3Large, extracting meaningful visual features before being fed into the sequential model. This frame-wise analysis enabled the system to capture subtle variations in motion and detect patterns indicative of shoplifting behavior, even in cases where actions were brief or partially obscured.

Model Training with Recurrent Neural Networks

For the classification task, we implemented a recurrent neural network (RNN) architecture with long short-term memory (LSTM) units. The LSTM-based network was designed to process the sequential frame data extracted from MobileNetV3Large, capturing temporal relationships that distinguish shoplifting actions from normal customer behavior. The training process involved optimizing the network's weights using a loss function suitable for binary classification. The model was trained using the training dataset, with hyperparameter tuning performed to enhance performance. The final evaluation was conducted using the test dataset to measure the model's ability to generalize effectively.

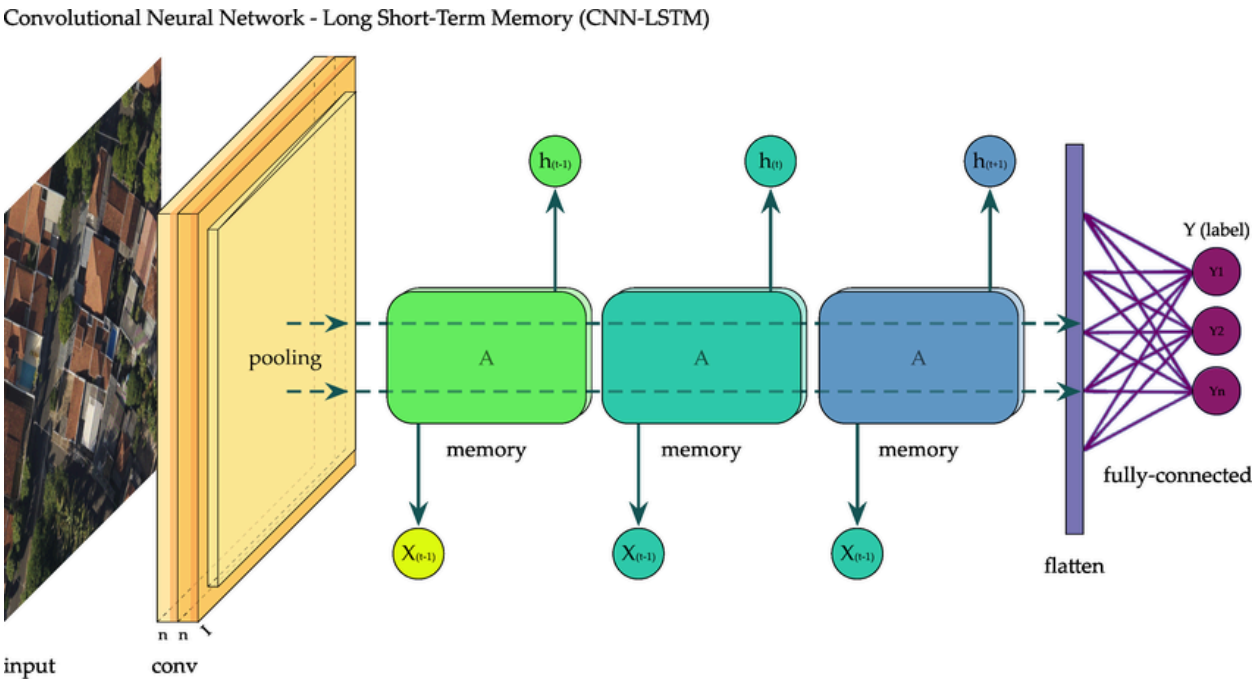

Figure 3. The main stages of the classification algorithm CNN-LSTM

YOLO-Based Object Detection

In addition to frame-level feature extraction, we integrated the YOLO (You Only Look Once) object detection algorithm to enhance shoplifting recognition. YOLO was employed to detect key objects such as shopping bags, clothing, and hand movements in real-time. This step was crucial for identifying interactions between individuals and objects that could indicate potential shoplifting behavior. By leveraging YOLO, we were able to filter and focus on critical areas within frames, significantly improving the model's ability to differentiate between normal customer behavior and suspicious activities.

Figure 2 provides a detailed visualization of the **experimental** setup, showcasing the key components and methodology employed in this study.

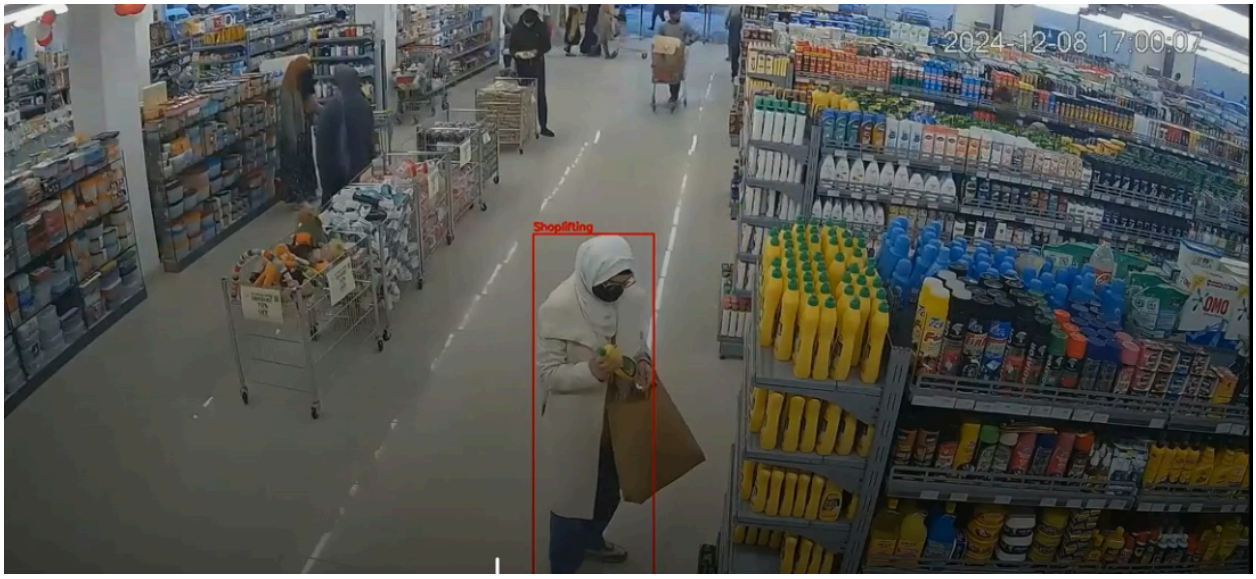

Figure 4. Visualization of our **experimental** setup

The program for solving the problem of video surveillance shoplifting recognition was written in the Python programming language. The web-based interactive computing environment, Jupyter Notebook, was used for dataset processing, while the Colaboratory environment from Google Research was utilized to design the neural network architecture and to train and test the model. This resource enabled the program to run in a cloud environment with Google Tensor Processing Units (TPUs), ensuring high-speed performance when working with machine learning frameworks, specifically the TensorFlow library, which performed tensor-based computations.

6. Findings and Results

.To achieve a sufficiently high accuracy of the classifier, extensive research and experimentation were conducted in practice. This included selecting an appropriate video classification method, finding a suitable dataset, optimizing data processing techniques, and configuring neural networks with optimal parameters. Data processing played a crucial role in preparing the model for training. As noted earlier, we performed data augmentation to artificially expand the dataset. Training on 310 samples resulted in an accuracy of approximately 77%, whereas increasing the dataset to 1860 samples improved accuracy by around 5–7%.

The choice and configuration of neural networks were also critical factors. Initial experiments were conducted using a combination of the convolutional InceptionV3 network and a recurrent neural network with gated nodes. However, the speed of real-time inference was low due to the computationally expensive feature extraction process. To address this, we sought a network with fewer output features while still effectively capturing essential information in each frame. Through multiple experiments, we found an optimal neural network: MobileNetV3Large, which

produced an output size of (1,960) compared to the significantly larger (1,2048) of InceptionV3. As a result, accuracy improved by approximately 10%.

Among the key hyperparameters of the convolutional neural network was the dimension of the video fragment, which had to align with the input layer parameters. We standardized all frames to 224×224×3, maintaining an RGB color format. The top layer was removed for feature extraction, yielding a constant feature vector of size 1×960 per frame. Through trial and error, the optimal number of frames per sequence was determined to be 30.

Another crucial optimization step was tuning the batch size parameter, which defines the number of samples processed per training iteration. Based on our experimental results, we evaluated batch size of 64, analyzing their impact on training and validation accuracy. The dataset was split into training (70%) and validation (30%) sets, with the validation set serving as an intermediate step for model selection and optimization. As reflected in our results, selecting an appropriate batch size had a direct influence on model convergence and overall accuracy.

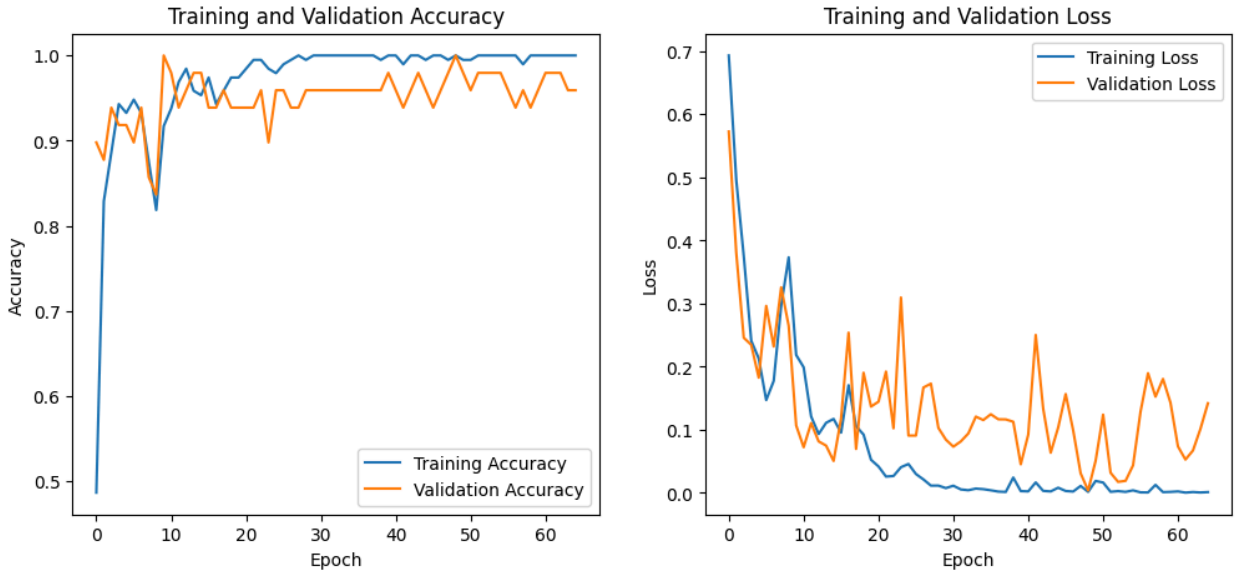

Figure 5. Training and validation accuracy and loss depending on the epoch

The confusion matrix presented provides a comprehensive evaluation of the model's performance in shoplifting recognition. The matrix indicates that the model correctly classified 25 instances of normal behavior and 22 instances of shoplifting, demonstrating a high level of accuracy. Notably, the false negative rate is zero, ensuring that all shoplifting cases are successfully identified, which is crucial for real-world surveillance applications. The model exhibits a low false positive rate, with only two instances of normal behavior being misclassified as shoplifting. This balance between precision and recall highlights the model’s effectiveness in distinguishing between normal and suspicious activities, making it a reliable solution for automated surveillance systems.

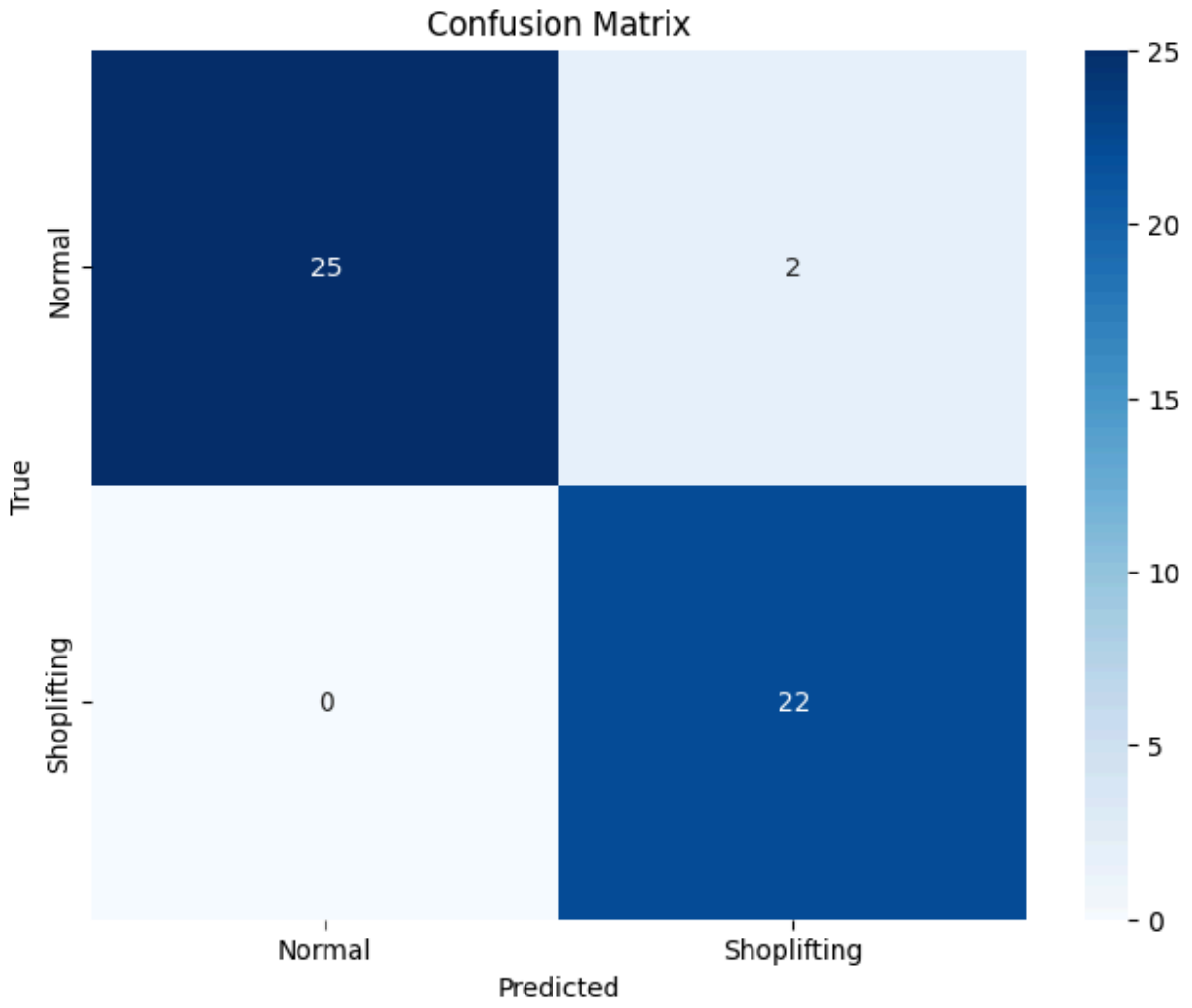

Figure 6. Confusion Matrix for Shoplifting Recognition Model

7. Conclusion

In this article, we propose a video classification model designed to detect shoplifting incidents in surveillance footage by integrating **ResNet34**, **YOLO**, and a **frame-level CNN**. Our approach combines spatial feature extraction and object detection to improve classification accuracy and robustness. Unlike traditional methods that rely on recurrent architectures, we employ a **CNN-RNN framework** to efficiently process both spatial and temporal information.

For spatial feature extraction, we utilize **ResNet34**, a deep learning model renowned for its residual connections, which facilitate the training of deeper networks without encountering vanishing gradient issues. Additionally, **YOLO (You Only Look Once)** is incorporated for real-time object detection, enabling the model to identify and focus on critical regions within

video frames. A **frame-level CNN** is then applied to refine the extracted features before feeding them into the recurrent network for temporal sequence analysis.

We trained and evaluated our model on a **custom dataset** specifically curated for shoplifting detection. This dataset ensures balanced class distribution and enhances the model's ability to generalize to real-world surveillance scenarios. By leveraging **transfer learning**, we utilized pretrained models to significantly reduce training time while maintaining high classification accuracy.

Our experimental results demonstrate that the proposed **ResNet34 + YOLO + frame-level CNN** framework effectively captures both spatial and temporal dependencies, achieving a classification **accuracy of 93%.** This performance highlights the model's suitability for real-time surveillance applications. In future work, we aim to optimize inference speed, refine object detection for more precise feature extraction, and explore transformer-based architectures to improve long-range temporal modeling. Further research will focus on the practical implementation of the proposed hybrid neural network in shopping malls

References

1. Chemere, D.S. *Real-time Shoplifting Detection from Surveillance Video.* Master Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2018; p. 94.
2. Kirichenko, L.; Radivilova, T. Analyzes of the distributed system load with multifractal input data flows. *In Proceedings of the 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM 2017)*, Lviv, Ukraine, 21–25 February 2017; pp. 260–264.
3. Gim, U.J.; Lee, J.J.; Kim, J.H.; Park, Y.H.; Nasridinov, A. An Automatic Shoplifting Detection from Surveillance Videos. *In Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 7–12 February 2020; Apress: Berkeley, CA, USA, 2020; Volume 34, pp. 13795–13796.
4. Ivanisenko, I.; Kirichenko, L.; Radivilova, T. Investigation of multifractal properties of additive data stream. *In Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing*, Lviv, Ukraine, 23–27 August 2016; pp. 305–308.
5. Kirichenko, L.; Radivilova, T.; Bulakh, V. Machine learning in classification time series with fractal properties. *Data* 2019, **4**, 5. [CrossRef]
6. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* 2020, **1**, 36. [CrossRef]
7. Radivilova, T.; Kirichenko, L.; Ageiev, D.; Bulakh, V. Classification methods of machine learning to detect DDoS attacks. *In Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Metz, France, 18–21 September 2019; pp. 207–210.
8. Rehman, A.; Belhaouari, S.B. Deep Learning for Video Classification: A Review. *TechRxiv* 2021, preprint.

9. Yamato, Y.; Fukumoto, Y.; Kumazaki, H. Security camera movie and ERP data matching system to prevent theft. *In Proceedings of the 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, 8–11 January 2017; pp. 1014–1015.
10. Tsushita, H.; Zin, T.T. A Study on Detection of Abnormal Behavior by a Surveillance Camera Image. *In Big Data Analysis and Deep Learning Applications*; Zin, T.T., Lin, J.C.W., Eds.; Springer: Singapore, 2019; pp. 284–291.
11. Flores-Munguia, C.; Ortiz-Bayliss, J.C.; Terashima-Marin, H. Leveraging a Neuroevolutionary Approach for Classifying Violent Behavior in Video. *Comput. Intell. Neurosci.* 2022, **2022**, 1279945. [CrossRef] [PubMed]
12. Morales, G.; Salazar-Reque, I.; Telles, J.; Diaz, D. Detecting violent robberies in CCTV videos using deep learning. *IFIP Advances in Information and Communication Technology*. *In Artificial Intelligence Applications and Innovations*; Springer International Publishing: Cham, Switzerland, 2019; pp. 282–291.
13. Akbar, M.N.; Riaz, F.; Awan, A.B.; Khan, M.A.; Tariq, U.; Rehman, S. A Hybrid Duo-Deep Learning and Best Features Based Framework for Action Recognition. *Comput. Mater. Contin.* 2022, **73**, 2555–2576.
14. Nasaruddin, N.; Muchtar, K.; Afdhal, A.; Dwiyantoro, A.P.J. Deep anomaly detection through visual attention in surveillance videos. *Big Data* 2020, **7**, 87. [CrossRef]
15. University of Central Florida. UCF-Crime Dataset. Available online: https://www.v7labs.com/open-datasets/ucf-crime-dataset (accessed on 17 September 2022).
16. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
17. Arunnehru, J.; Chamundeeswari, G.; Bharathi, S.P. Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos. *Procedia Comput. Sci.* 2018, **133**, 471–477. [CrossRef]
18. Li, J.; Jiang, X.; Sun, T.; Xu, K. Efficient Violence Detection Using 3D Convolutional Neural Networks. *In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
19. Islam, M.S.; Sultana, S.; Kumar Roy, U.; Al Mahmud, J. A review on video classification with methods, findings, performance, challenges, limitations, and future work. *J. Ilm. Tek. Elektro Komput. Dan Inform.* (JITEKI) 2020, **6**, 47–57. [CrossRef]
20. Alfaifi, R.; Artoli, A.M. Human action prediction with 3D-CNN. *SN Comput. Sci.* 2020, **1**, 286. [CrossRef]
21. Martinez-Mascorro, G.A.; Abreu-Pederzini, J.R.; Ortiz-Bayliss, J.C.; Garcia-Collantes, A.; Terashima-Marin, H. Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. *Computation* 2021, **9**, 24. [CrossRef]
22. Ansari, M.A.; Singh, D.K. ESAR, An Expert Shoplifting Activity Recognition System. *Cybern. Inf. Technol.* 2022, **22**, 190–200. [CrossRef]
23. Harvey, M. Five Video Classification Methods Implemented in Keras and TensorFlow: Exploring the UCF101 Video Action Dataset. 2017. Available online:

https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5 (accessed on 14 September 2022).
24. Kirichenko, L.; Alghawli, A.S.A.; Radivilova, T. "Generalized approach to analysis of multifractal properties from short time series." *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, pp. 183–198, 2020. [CrossRef]
25. J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 27–30 June 2016, pp. 2285–2294.
26. S. Gollapudi, *Learn Computer Vision Using OpenCV: With Deep Learning CNNs and RNNs*, 1st ed. Berkeley, CA, USA: Apress, 2019, p. 171.
27. C. Nebauer, "Evaluation of convolutional neural networks for visual recognition," *IEEE Trans. Neural Netw.*, vol. 9, p. 685, 1998. [CrossRef] [PubMed]
28. L. Medsker and L. C. Jain, Eds., *Recurrent Neural Networks: Design and Applications (International Series on Computational Intelligence)*, 1st ed. Boca Raton, FL, USA: CRC Press, 1999, p. 416.
29. "Time Series Classification. Welcome to the UEA & UCR Time Series Classification Repository." Available online: http://www.timeseriesclassification.com (accessed on 9 May 2022).
30. R. S. Segall and G. Niu, *Biomedical and Business Applications Using Artificial Neural Networks and Machine Learning*. Hershey, PA, USA: IGI Global, 2022, p. 394.
31. S. Shah, *Implementation and Evaluation of Gated Recurrent Unit for Speech Separation and Speech Enhancement*. DeKalb, IL, USA: Northern Illinois University, 2019, p. 91.
32. LazyProgrammer, *Deep Learning: Recurrent Neural Networks in Python: LSTM, GRU, and More RNN Machine Learning Architectures in Python and Theano*. Berkeley, CA, USA: Apress, 2021, p. 93.
33. S. A. Medjahed, "A Comparative Study of Feature Extraction Methods in Images Classification," *Int. J. Image Graph. Signal Process.*, vol. 7, p. 16, 2015. [CrossRef]
34. "ImageNet Database." Available online: https://image-net.org/index.php (accessed on 8 July 2020).
35. "Keras API Reference/Keras Applications/MobileNet, MobileNetV2, and MobileNetV3." Available online: https://keras.io/api/applications/ (accessed on 14 September 2022).