International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Anomaly Detection in CCTV Videos Using LSTM Architecture

1st Agashini V Kumar
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India

2nd Suresh D
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs081@jainuniversity.ac.in

3rd Saurabh Rai
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs069@jainuniversity.ac.in

4th Ritik Raj
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs059@jainuniversity.ac.in

5th Rohit Raj
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs060@jainuniversity.ac.in

Abstract: The need for effective, automated anomaly detection systems that can handle enormous volumes of video data has been highlighted by the sharp rise in CCTV surveillance in both public and private areas. In order to improve situational awareness and give security personnel useful intelligence, this study proposes a model based on deep learning for identifying and measuring anomalies in CCTV data. Utilizing computer vision, the Convolutional neural networks (CNNs) are used in the model to extract spatial information. which allows it to identify and categorize anomalous occurrences in a variety of contexts. Furthermore, Long Short-Term Memory (LSTM) networks or recurrent neural networks (RNNs) are utilized to investigate mobility and activity patterns across time. This enhances the model's comprehension of sequence-based activities and enables it to identify deviations across time. Transfer learning techniques are used to improve performance even more, enabling the model to adapt effectively in a variety of settings without requiring a lot of retraining. By determining a numerical threshold score based on the anomaly's attributes, including frequency, intensity, and kind, this method is novel in that it can not only identify anomalies but also evaluate their seriousness. Because it enables security teams to rank answers based on the evaluated threat level, this severity score is essential for more efficient resource allocation and quicker reaction times.

I. Introduction

A cutting-edge technique for improving security systems is the use of Computer vision and deep learning to detect anomalies in CCTV data. This project's objective is to develop an automated system that can identify anomalous activity in real time. The system will examine CCTV camera video feeds and identify departures from typical patterns of activity by utilizing developments in computer vision and deep learning. Such technology can enhance response times in emergency circumstances, boost the effectiveness of surveillance systems, and drastically lessen reliance on human operators. By creating an automated anomaly detection system Using computer vision and deep learning technology, this study aims to address the limitations of the surveillance systems in place today. In this context, anomalies are defined as odd patterns or behaviors that are out of the ordinary and may be signs of possible security risks. We present a deep learning algorithm that can recognize any unusual or suspicious activity in real-time CCTV data. In order to train our model with a variety of datasets, it is crucial to record both live CCTV footage and CCTV film that has previously displayed suspicious activity.

Our goal is to create a model that can automatically identify undesirable behavior and human nature

from live CCTV footage and notify the appropriate authorities. Metal detectors can be replaced with contemporary deep learning techniques to identify individuals carrying undesired weapons and notify authorities of any unexpected activity approaching a restricted area or authority. To spot irregularities instantly, the system will make use of deep learning algorithms, which are especially made to examine and comprehend video data. To improve the system's comprehension and interpretation of intricate visual data from CCTV footage, computer vision techniques will be incorporated. The system's automation of anomaly detection will increase overall security by facilitating speedier identification and reaction to possible threats, in addition to increasing the effectiveness of monitoring operations.

II. Related Works

Because anomalous events are unpredictable and visual cue interpretation is difficult, researchers have proposed a wide range of models for deep learning to identify and localize anomalies in video sequences. Anomaly detection in CCTV surveillance footage has become a crucial area of study within computer vision and deep learning, with the goal of ensuring public safety through automated video monitoring.

In this field, convolutional neural networks, or CNNs, continue to be fundamental. The extraction of spatial characteristics from frames has been made easier by architectures such as ResNet and Mask R-CNN, which have greatly improved instance-level segmentation and object detection [1][2]. Mask R-CNN is a useful tool in surveillance applications because of its adaptability and ease of use, which enable it to do tasks like pose estimation in addition to object recognition [1]. By suggesting architecture changes centered on speed and processing efficiency—two critical characteristics for deployment in real-time CCTV systems with limited resources—ShuffleNet V2 improved CNN design even more [3].

Hybrid models and lightweight networks tailored to particular surveillance tasks have gained more interest in recent years. For example, SimpleNet shows good accuracy and efficiency in real-world scenarios by combining an anomaly discriminator and a pre-trained feature extractor [4]. Another development, PASS-CCTV, demonstrates the resilience necessary for real-world implementation by addressing anomaly detection under challenging environmental circumstances such as poor sight or severe weather [5].

Given that surveillance data sometimes consists of lengthy, continuous video streams, temporal modeling is equally important. Although more recent approaches currently prioritize attention mechanisms and transformers for better performance in modeling sequence-level data, long short-term memory (LSTM) models, and recurrent neural networks (RNNs) have been used to capture temporal connections between frames. A new class of transformer-inspired models, such as PromptAD, are very effective in situations with little aberrant data because they use rapid engineering to utilize few-shot learning [9].

Unsupervised models and autoencoders also make substantial contributions to the discipline. A new memory-based unsupervised method called SoftPatch denoises data at the patch level, hence preventing overfitting caused by label noise, which is a prevalent problem in real-world surveillance datasets [6]. In a similar vein, PatchCore achieves good AUROC scores on datasets such as MVTec AD [11] by using representative memory banks to differentiate between nominal and aberrant activity.

The practical integration of these methodologies is demonstrated by a number of applied studies. For example, AIGuard achieves over 85% recognition accuracy for frontal faces by using Multi-task Cascaded Convolutional Networks (MTCNN) for face detection and a deep ResNet model to track known individuals in surveillance footage [7]. A-eye automates attendance tracking from low-resolution CCTV feeds in school settings using CNNs trained on student face databases, demonstrating 76% recognition accuracy even in a

variety of illumination and orientation circumstances [8].

Domain-specific modifications of these methods are examined in other recent studies. To improve anomaly recognition at night or in dimly lit environments, Sandhiya et al., for instance, established a framework that uses contrast enhancement and HDR imaging to improve low-light CCTV footage [12]. The need for latency-aware, fast inference is highlighted by real-time systems, such as those created by Poonia et al., which use deep models to identify crimes in live broadcasts [10].

When taken as a whole, these contributions offer a comprehensive method of anomaly detection that incorporates developments in network architecture, feature representation, deployment robustness, and temporal modeling. In addition to being more accurate, the discipline is working to develop models that are scalable, interpretable, and resistant to the dynamic, noisy nature of real-world CCTV data.

III. Methodology

A. Data Collection

The benchmark anomaly detection dataset, first presented by Sultani et al. [13], served as the basis for the dataset utilized in this investigation. For anomaly detection jobs in actual surveillance scenarios, it has been further improved and given a label. 13 distinct anomaly classes— The collection includes representations of abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, theft, shoplifting, and vandalism. , which consists of 16,853 video clips. Depending on whether abnormal activity is present, each Both normal (0) and pathological (1) labels are applied to the video.

A balanced and varied collection of real-world events is provided by the 9,676 movies that are classified as normal and 7,177 as abnormal, which enable efficient training and assessment of anomaly detection algorithms.

Each video was processed by extracting individual frames at a consistent frame rate to make model training easier. In order to preserve uniformity throughout the collection, these frames were shrunk to a fixed resolution. Pixel values between 0 and 1 were scaled using standard normalization procedures. The dataset enables both supervised and semi-supervised learning techniques because it already includes labeled data. In order to maintain the authenticity of actual surveillance situations, no artificial augmentation techniques were used in this study.

B. Dataset Preprocessing and Clip-wise Frame Extraction

The raw dataset needed extensive preparation to standardize its structure and get it ready for effective data loading in order to make training an anomaly detection model easier. The surveillance video clips in the dataset were divided into various anomaly types, and each category included many full-length video files that were then divided into shorter sub-clips. Uncertainty during data access resulted from the labels' reference to sub-clip identifiers rather than a straightforward mapping to the files' actual locations within the directory hierarchy.

Aligning the label information with the dataset's physical arrangement was the first step in the preprocessing stage. Reconstructing the proper hierarchical pathways that led to the associated sub-clips within their respective parent video folders required deciphering each clip's naming convention. In order to remove discrepancies between the label files and the dataset structure, a new mapping was created to reflect the full and precise file locations for every labeled clip. After the alignment, all tagged sub-clips and associated anomaly classifications were combined into a comprehensive annotation file. During training and assessment, each video clip may be programmatically retrieved and linked to its appropriate label thanks to this file's dependable reference for further data processing stages.

The extraction of individual frames from each sub-clip was the following step. Using frame sampling techniques, each video clip was broken down into a

series of image frames. These frames were kept in specific folders that were arranged based on the names of the respective clips. In order to enable frame-wise analysis, which is frequently necessary for real-time anomaly detection applications, this step was crucial. By preventing recurrent video decoding, pre-extracted frames greatly decreased the computational overhead during training. The output of this one-time preparation workflow was a frame-level dataset with accurate annotations and a well-structured structure. The development of an efficient anomaly detection system was made possible by this conversion of the unprocessed video data into a format that was compatible with the model.

C. Model Architecture

In this work, we employ a deep learning-based sequence model to identify irregularities in CCTV footage, which includes Long Short-Term Memory (LSTM) layers [14]. Utilizing the temporal dynamics of video sequences to differentiate between typical and abnormal activity is the main concept [15]. Using a Convolutional Neural Network (CNN) that has already been trained, like I3D, EfficientNet, or a comparable backbone, which efficiently captures spatial attributes, each video clip is first processed to extract frame-level information [16]. The LSTM-based model is then fed these features [17].

Two stacked LSTM layers make up the architecture, which is followed by dense layers for classification and dropout regularization [18]. The purpose of Contextual knowledge and temporal relationships between video frame sequences are learned by the LSTM layers. The model's Capacity to record temporal characteristics at both low and high levels—both essential for spotting minute abnormalities across time—is achieved by stacking numerous LSTM layers [19]. The nature of surveillance recordings, where the temporal context is frequently more suggestive of anomalies than individual frames, makes this approach especially well-suited for anomaly detection tasks [20]. An individual frame, for example, can seem normal, but when viewed in a sequence, it might indicate an action that is suspicious or inappropriate.

Based on the expected label distribution, the model's final dense layers, which include a softmax output, enable it to produce class probabilities for preset activity categories, which can be used to deduce normal or abnormal behavior [21].

The necessity for a computationally efficient yet temporally sensitive model that works well with sequences of extracted features rather than raw video data was the main factor in the choice of LSTM over other architectures like 3D-CNN or Transformers. The ability to handle variable-length sequences and learn long-term dependencies without consuming a lot of compute power is another benefit of LSTMs.

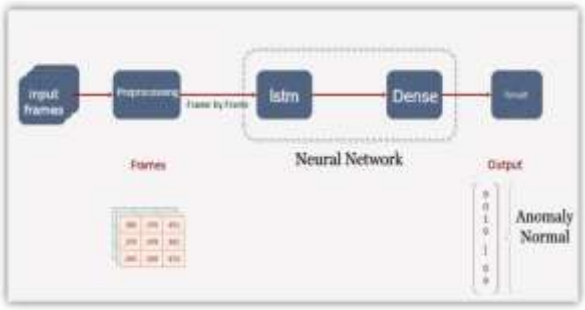

Figure 1 Model Architecture

D. Model Training and Validation

We used the following training and validation methodology to guarantee reliable performance estimation and reproducibility:

1. Data Splitting

All frames were first divided using a fixed random seed (random_state=42) into training (80%) and test (20%) sections.to guarantee consistent splits across runs. During model fitting, an additional 20% of the training data was reserved for validation by specifying validation_split=0.2.

2. Hyperparameter Settings

Model was trained with the **Adam optimizer** for **25 epochs** and a **batch size of 36**. at its default learning rate of 1×10^{-4}. Performance on the validation set was monitored after each epoch, the model weights are evaluated on a held-out validation set (15 % of the data). We save the checkpoint with the lowest validation loss and roll back to it at the end of training to guard against overfitting.

3. Regularization

Dropout: Following the LSTM layers, a dropout rate of 0.4 is applied toreduce co-adaptation of neuronss.

Weight decay: L2 regularization (1×10^{-5}) is applied to all trainable weights in the dense layers.

E. Evaluation Metrics

Accuracy and loss were the two main measures used to assess The efficiency of the suggested anomaly detection algorithm. These indicators show how successfully the model adapts to fresh data and how well it recognizes patterns in the training set.

The percentage of accurately anticipated cases relative to all predictions is known as accuracy. It offers a comprehensive comprehension of the model's functionality and is a commonly utilized statistic for classification jobs. The model's capacity to accurately categorize video sequences as either normal or aberrant is reflected in this work. The model is producing more accurate predictions as its accuracy is higher.

Conversely, loss is a measurement of the model's prediction inaccuracy. The model's goal during training is to reduce the loss function, which is categorical cross-entropy loss in this instance. The disparity between the actual class labels and the expected probability distribution is measured by this function. Better learning is indicated by a smaller loss value, which indicates that the model's projected probabilities are closer to the true labels. Finding problems like overfitting or underfitting is made easier by tracking the loss across training and validation datasets.

IV. Results and Discussions

The performance of the recommended approach in identifying harmful online comments is shown and discussed in this chapter. Accuracy and loss measures datasets for training and validation were utilized to monitor the model's performance during the 25 epochs of training and evaluation. Figures 2 and 3 provide a graphic depiction of the training progress.

A. Model Accuracy Analysis

The accuracy of training and validation during 25 epochs is shown in Figure 2. After the first epoch, the model's initial training accuracy was roughly 52.5%, and its validation accuracy was roughly 64.3%. Both accuracies steadily improved with each new era. By epoch 25, the validation accuracy was 85.1% and the training accuracy was 92.2%. This upward trend suggests that the model successfully picked up patterns from the training set. Although it fluctuated occasionally, the validation accuracy also generally increased. Although the performance is still strong overall, the discrepancy between training and validation accuracy in subsequent epochs may indicate slight overfitting.

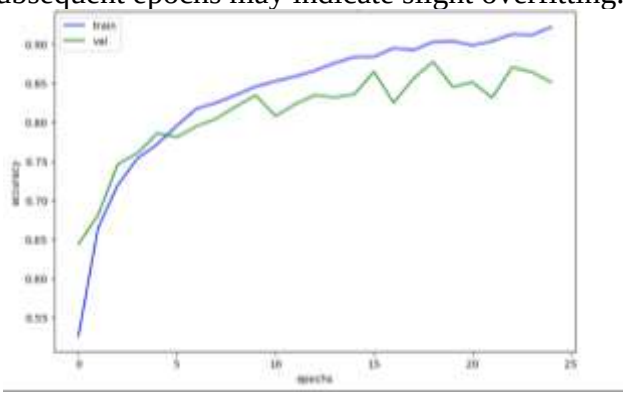

Figure 2 Model Accuracy

B. Model Loss Analysis

The model's Figure 3 displays the loss for both the training and validation sets. The training loss was 1.4151 at the beginning, but it dropped significantly in the first few epochs and then progressively until it reached 0.2204 by epoch 25. A similar pattern was seen in the validation loss, which began at 1.0944 and decreased to 0.4908 by the conclusion of the epoch.

Epoch 17 showed the lowest validation loss of

0.3739, indicating that this was the region where the model generalized the best. Though there are some indications of overfitting in subsequent epochs as a result of the growing disparity between training and validation losses, the final validation loss still shows good generalization.

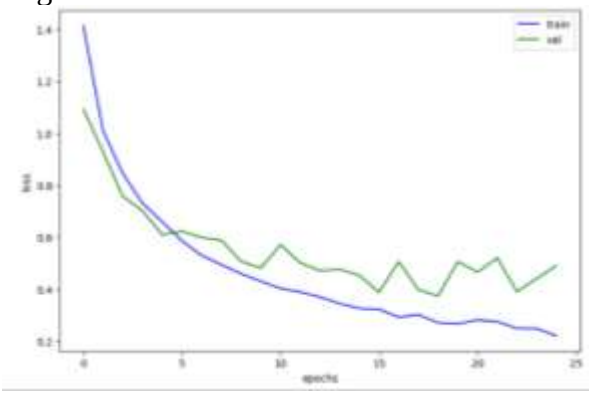

Figure 3 Model Loss

V. Conclusion

In this study, we created and assessed a deep learning-based sequence model for the categorization of harmful internet comments. The model learned temporal dependencies and patterns among sequences of characteristics collected from comment text by using an LSTM-based architecture. For reliable categorization into hazardous categories, the architecture included two stacked LSTM layers, followed by dense layers and dropout regularization.
According to our experimental findings, which are covered in Chapter 4, the model demonstrated a great capacity for learning and generalization by reaching the 25th epoch with a high training accuracy of 92.2% and a validation accuracy of 85.1% . Convergence efficiency was further improved during training by combining the Adam optimizer with categorical crossentropy as the loss function.
The accuracy and loss graphs (Figures 2 and 3) show trends that support the model's ability to identify harmful content. Even though there was some overfitting in the later periods, the The

model's consistent performance indicates that it is prepared for deployment in real-world moderation systems.

To further increase classification accuracy, this model can be expanded in the future with features like transformer-based encoders, hybrid CNN-LSTM architectures, or attention techniques. Scalable deployment for real-time content moderation on social media and discussion platforms can also be ensured by integration with DevOps processes.
All things considered, this study offers a useful method for addressing the continuous problem of automating online toxicity detection by providing a deep learning solution that is both efficient and flexible.

Referances

[1] He, K., G. Gkioxari, P. Dollár, and R. B. Girshick (2017). Mask R-CNN.

[2] Cai, Zhao Wei, and Nuno Vasconcelos. "Cascade R-CNN: Delving Into High Quality Object Detection." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017): 6154–6162.

[3] Ma, Ningning, et al. "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design." ArXiv abs/1807.11164 (2018): no pagination.

[4] Liu, Zhikang et al. "SimpleNet: A Simple Network for Image Anomaly Detection and Localization." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023 (20402-20411).

[5] Jeon, Hobeom et al. "PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions." Expert System Application 254 (2024): 124391..

[6] Jiang, Xi et al. “SoftPatch: Unsupervised Anomaly Detection with Noisy Data.” *ArXiv* abs/2403.14233 (2024): n. pag.

[7] Singh, Aryan and Rohit Kumar Tiwari contributed to "AIGuard: Criminal Tracking in CCTV Footage Using MTCNN and ResNet." 2024

14th International Conference on Cloud Computing, Data Science, and Engineering (Confluence): 31-35.

[8] Ambre, Nimesh et al. "A-eye: Attendance monitoring using face detection and recognition from CCTV footage." *2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA)* (2024): 1-6.

[9] Li, Xiaofan et al. "PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024): 16848-16858.

[10] Poonia, Ramesh Chandra et al. "Finding Real-Time Crime Detections During Video Surveillance by Live CCTV Streaming Using the Deep Learning Models." *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence* (2024): n. pag.

[11] Roth, Karsten et al. "Towards Total Recall in Industrial Anomaly Detection." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 14298-14308.

[12] Sandhiya, B. et al. "Crime Investigation System using CCTV Footage: A Novel Framework for Contrast Enhancement of Dark Images." *2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)* (2024): 462-467.

[13] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Center for Research in Computer Vision (CRCV)*, 1–10.

[14] Ul Amin, Sareer et al. "Video Anomaly Detection Utilizing Efficient Spatiotemporal Feature Fusion with 3D Convolutions and Long Short-Term Memory Modules." *Adv. Intell. Syst.* 6 (2024): n. pag.

[15] Tang, Yiyin et al. "Semi-supervised LSTM with historical feature fusion attention for temporal sequence dynamic modeling in industrial processes." *Eng. Appl. Artif. Intell.* 117 (2023): 105547.

[16] Ullah, Waseem et al. "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks." *Multimedia Tools and Applications* 80 (2020): 16979 - 16995.

[17] AlMahadin, Rawiah, et al. *Enhancing Video Anomaly Detection Using Spatio-Temporal Autoencoders and Convolutional LSTM Networks*. 2024.

[18] Wei, Xinyu, et al. "Detecting Video Anomaly with a Stacked Convolutional LSTM Framework." *Pattern Recognition Letters*, vol. 125, 2019, pp. 621–627.

[19] Ullah, Ahsan, et al. "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos." *IEEE Access*, vol. 9, 2021, pp. 128116–128129.

[20] Cheng, Gang, et al. "An Anomaly Comprehension Neural Network for Surveillance Videos on Terminal Devices." *IEEE Transactions on Multimedia*, vol. 22, no. 10, 2020, pp. 2674–2685.

[21] Amin, M. Usama, et al. "EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos." *Applied Soft Computing*, vol. 113, 2022, 107939.

[22] Durrant, Thomas, et al. "3D Convolutional and Recurrent Neural Networks for Reactor Perturbation Unfolding and Anomaly Detection." *Annals of Nuclear Energy*, vol. 132, 2019, pp. 420–431.

[23] Luo, Wen, et al. "Remembering History with Convolutional LSTM for Anomaly Detection." *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 439–444.