Mono-scale CNN-LSTM fusion network for Suspicious Activity Detection

AASMA AAS

S2022279005

DR. SYED FAROOQ

SCHOOL OF SYSTEMS AND TECHNOLOGY

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY
LAHORE, PAKISTAN

2022 -2024

AASMA AAS

S20222279005

DR. SYED FAROOQ ALI (SUPERVISOR)
KHALID IJAZ (CO-SUPERVISOR)

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

SCHOOL OF SYSTEMS AND TECHNOLOGY
UNIVERSITY OF MANAGEMENT AND TECHNOLOGY,
LAHORE, PAKISTAN

2022-2024

UNIVERSITY OF MANGEMENT AND TECHNOLOGY

ORIGINAL LITERARY WORK DECLARATION

Name of Student: **AASMA AAS**

Registration No: **S2022279005**

Name of Degree: **Master of Science in Computer Science**

Title of Research Report/Dissertation/Thesis:

Mono-scale CNN-LSTM fusion network for Suspicious Activity Detection

Field of Study: **Computer Vision**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any usage of any work in which copyright exists was done by strategy for sensible overseeing and for permitted purposes and any bit or concentrate from, or reference to or expansion of any copyright work has been uncovered expressly and enough and the title of the Work and its introduction have been perceived in this Work;
(4) I don't have any genuine learning nor do I ought to reasonably to understand that the creation of this work contains an infringement of any copyright work;
(5) I subsequently dole out all and each straightforwardly in the copyright to this Work to the University of Management and Technology ("UMT"), who starting now and into the foreseeable future will be owner of the copyright in this Work and that any multiplication or use in any structure or utilizing any and whatsoever methods all is confined without the made consent out of UMT having been first had and gained;
(6) I am completely attentive that if over the extent of influencing this Work I to have infringed any copyright whether deliberately or fiasco will approach, I may be in peril to substantial action or some other advancement as may be regulated by UMT.

Student's Signature: _____ Date: _____

CERTIFICATE OF APPROVAL

It is certified that the research work presented in this thesis entitled

"Mono-scale CNN-LSTM fusion network for Suspicious Activity Detection"

was conducted by

AASMA AAS, ID: S2022279005
under the supervision of
DR. SYED FAROOQ ALI
No part of this thesis has been submitted anywhere for any other degree.
This thesis is submitted to
Department of ***SYSTEM AND TECHNOLOGY,***
University of Management and Technology
for the partial fulfillment of the requirement for the degree of
Master of Science in *Computer Science*
at the Department of *System and Technology*,
University of Management and Technology, Lahore, Pakistan.

_____	_____
Supervisor	Signature
_____	_____
Chairperson	Signature
_____	_____
Dean SST	Signature

ACKNOWLEDGEMENTS

All praise to Almighty Allah, most accepting and most humane who enabled me and has given me capacities to do some research work and to add to the honorable field of learning. This thesis depicts the examination work embraced at School of Systems and Technology, University of Management and Technology, Lahore under the supervision of **Dr. Syed Farooq Ali** to whom I am very obligated for supervision, proposing the subject, entire time direction, support, and recommendations. His recommendations, discussions, directions, and remarks were consistently a source of inspiration and enthusiasm for me. I would like to express my special gratefulness and thanks to **Khalid Ijaz**. I would like to thank you for encouraging me in my research and for allowing me to grow. Your advice, counsels, and inspirations have always been priceless in every field of my life. I want to thank you for providing Peer review of my research work and help me identify my mistakes.

Last but not least, I feel inadequacy in vocabulary to discover appropriate words to precise my emotions for my Parents who raised up and groomed me all through of my life, whose hands are constantly raised for prayers which made me active in each field of my life. Their day and night support for me empowers me to join higher thoughts of life, taking care of all of the issues and to achieve my goals.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Fully Autonomous video surveillance system is an extensive and broad research area within the domain of computer vision. In recent studies, multiple approaches have been explored to detect crime activities. However, the mentioned approaches require manual monitoring that is time-consuming. To address this issue this thesis presents a state-of-the-art deep learning approach for autonomous video surveillance that focuses on highly accurate detection of suspicious activities. The proposed model Mono-scale CNN-LSTM model, where Oriented FAST and Rotated BRIEF (ORB) technique is used for feature extraction and then combined CNN with LSTM for the recognition task. To validate the effectiveness of the CNN-LSTM model, extensive experiments were conducted, comparing its performance against several pre-trained deep learning models such as CNN, VGG-16, VGG-19, ResNet-50, and DenseNet etc. The experiments were conducted using the UCF crime image dataset, which contains a diverse set of scenarios representing various criminal activities such as Robbery, Fighting and Shoplifting. The proposed approach CNN-LSTM model demonstrated a remarkable improvement in accuracy, achieving an impressive 99% accuracy rate in detecting suspicious activities. This accuracy far surpasses that of other deep learning models commonly used in video surveillance.

CHAPTER 1: INTRODUCTION

Surveillance videos have become an increasingly active research area in image processing, deep learning, and computer vision nowadays. Derived from the Organized Crime Index [1], the majority of the global population (79.2%) resides in nations characterized by heightened levels of criminality (Figure1).

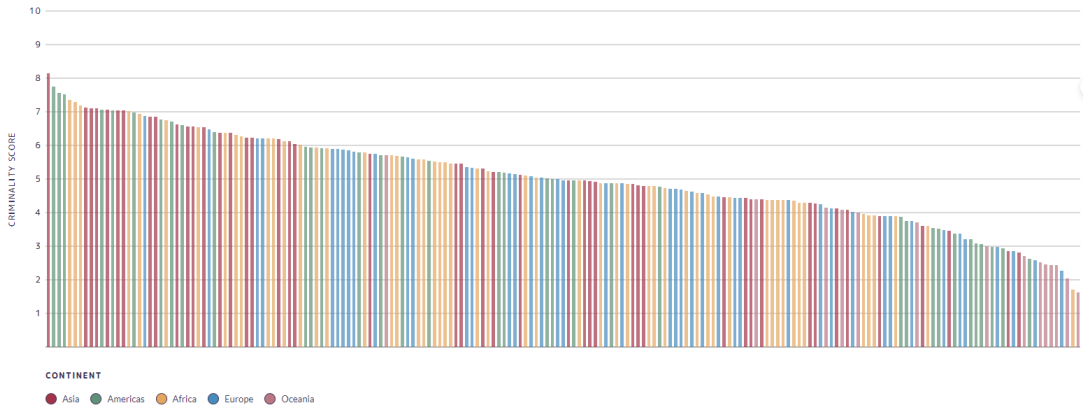

Figure 1: Globally Crime Index

Human activity recognition or suspect detection from video surveillance [27–30] has becoming popular in domain of image processing and computer vision research [26]. This thesis targets the crime problems such as Fighting, Shoplifting and Robbery. This research will help in creating a computerized system which will automatically detect crime activities and will alert individuals through alarm.

1.1 Problem Statement

The contemporary landscape is characterized by an increasing need for effective and efficient surveillance and security measures, driven by rising concerns over criminal activities. Among these, two prominent challenges are the detection and prevention

of robbery, fighting and shoplifting. These criminal acts not only threaten public safety but also exert economic and societal impacts.

Traditional surveillance methods typically rely on human monitoring, a labor-intensive and error-prone process susceptible to fatigue. Moreover, their effectiveness in identifying and responding to criminal activities is limited. Hence, there is a critical need to develop innovative and automated surveillance and detection systems that can consistently and accurately recognize suspicious behaviors associated with robbery, fighting and shoplifting. These systems should serve as a proactive deterrent while streamlining response efforts, contributing to the overall safety and security of communities and businesses. This research seeks to address these challenges by harnessing advanced deep learning and computer vision techniques to create intelligent surveillance systems capable of automating the detection and prevention of these crimes.

1.2 Motivation

This research is motivated by the imperative to address the challenges of surveillance and security systems in the detection and prevention of criminal activities like robbery, fighting and shoplifting. The limitations of human monitoring, marked by fatigue and inefficiency, underscore the need to automate surveillance processes. The advanced deep learning and computer vision technologies provides an opportunity to revolutionize video surveillance. Leveraging Convolutional Neural Networks (CNNs) for image processing and Long Short-Term Memory (LSTM) models for sequential data, the aim is to autonomously classify criminal activities with high accuracy. Furthermore, the motivation is fueled by the desire to benchmark the proposed deep

learning framework against established models, such as CNN, VGG-16, ResNet-50, and DenseNet etc., demonstrating its superiority [51]. Ultimately, this research seeks to contribute to enhanced public safety and security by pioneering an intelligent surveillance model that reduces dependence on human monitoring and leverages cutting-edge technologies for more effective crime prevention and response strategies.

1.3 Research Objectives

The primary research objective to alleviate the challenges associated with continuous human monitoring in surveillance systems. Monitoring activities with the human eye can be arduous and prone to errors due to factors like fatigue and human limitations. To address this, the research endeavors to develop a sophisticated surveillance model. This model will eliminate the need for constant human vigilance by incorporating an innovative alarm feature within the application. This feature will autonomously detect and trigger alerts when suspicious activities, such as robbery and shoplifting, are detected. By achieving this objective, the research aims to revolutionize surveillance by making it more automated and less reliant on human oversight.

Secondly, the central objective is to create and implement an advanced Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model. This model is designed to process video frames using a sequence of deep learning layers, including convolutional, activation, max-pooling, and dropout layers. The primary purpose of this model is to classify criminal activities based on probabilities accurately. Leveraging state-of-the-art deep learning techniques, this research objective seeks to

develop a robust and efficient system capable of precisely identifying and categorizing suspicious behaviors within surveillance footage.

In summary, the research objectives encompass automating surveillance processes, reducing the reliance on human monitoring, and developing a highly capable CNN-LSTM model for the accurate classification of criminal activities. These objectives collectively aim to enhance the effectiveness and efficiency of video surveillance systems, ultimately contributing to improved security and public safety.

1.3.1 Research Questions

- How an alarm feature generated to monitor the criminal activities by using proposed model.
- How can CNN model processes images through convolutional, activation, max-pooling, and dropout layers to classify criminal activities based on probabilities.
- How is the LSTM model handle sequential data in detecting criminal activities.
- How proposed deep learning framework is tested and validated on the UCF crime benchmark datasets, and how performance is compared with CNN, VGG-16, ResNet-50, and DenseNet models, etc.

1.4 Scope of Study

The core objective of this study revolves around the creation and deployment of a specialized Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model in the context of surveillance for robbery, fighting and shoplifting. This model is designed to process video frames by employing an array of deep learning layers, including convolutional layers for feature extraction, activation layers for non-linearity, max-pooling layers for spatial down sampling, and dropout layers for

regularization. The primary outcome sought is the accurate classification of suspicious behaviors associated with criminal activities.

1.5 Organization of the Thesis

The thesis remaining sections are structured as follows. The relevant literature on deep learning uses in the crime activities is presented in Section II. Section III describing the research problem background. In section VI, the experiments and results are presented whereas this study concluded in Section VII, which briefly discusses current and future research issues.

- **Chapter 2:** The chapter describes a survey of the literature related to the existing approaches and improvement of suspicious activity detection.
- **Chapter 3:** The chapter discusses the background of research target problem.
- **Chapter 4:** The chapter discusses the methodology of the proposed framework.
- **Chapter 5:** The chapter covers the findings and analyses.
- **Chapter 6:** The chapter concludes the thesis and future work is discussed.

CHAPTER 2: THEORETICAL BACKGROUND

2.1 Background

Researchers have embarked on an unceasing quest to innovate and refine algorithms and techniques for enhancing the accuracy and efficiency of detecting suspicious activities in surveillance videos. These advancements are not merely confined to academic curiosity, they hold the promise of vastly improving real-time monitoring and response systems, which are integral to upholding public safety. As we know that Traditional security systems (Figure 2) are time consuming and have biasness so the integration of Artificial Intelligence (AI), Deep Learning (DL), and Machine Learning (ML) has emerged as a transformative force in recognizing and classifying abnormal activities within the realm of surveillance videos.

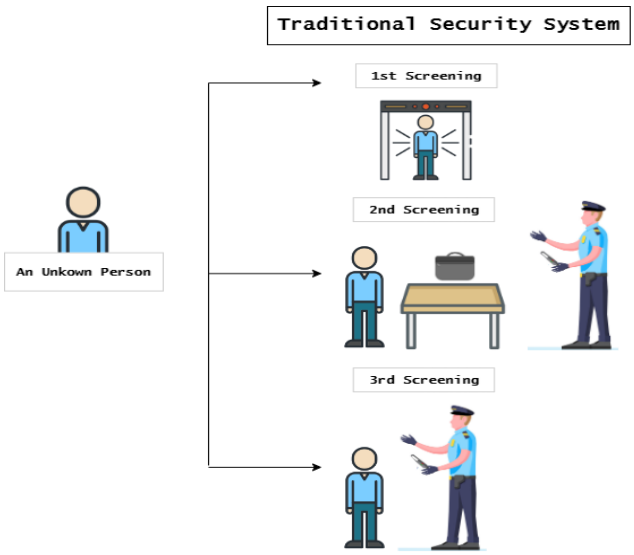

Figure 2: Existing Traditional Security System

At the heart of surveillance technology, there exist two distinct paradigms: the traditional surveillance system and the autonomous surveillance system. In the traditional approach, surveillance is primarily reliant on manual monitoring, which is resource-intensive and prone to human error. In stark contrast, the autonomous surveillance system is powered by the latest advancements in DL, AI, and ML. These autonomous systems have the capacity to autonomously train themselves to detect human activities with remarkable precision. They leverage their prowess in feature extraction and pattern recognition to excel in the detection of complex human behaviors.

One of the key strengths of deep learning models, which form the backbone of autonomous surveillance systems, lies in their ability to extract intricate representations of human actions through exposure to extensive datasets. These models, when adequately trained, demonstrate exceptional adaptability and generalization capabilities, even when faced with unseen data. This capacity is particularly vital in the dynamic realm of surveillance, where complex real-world scenarios, including occlusions, variable lighting conditions, and crowded environments, are the norm.

Nonetheless, the journey does not end with the development of deep learning models; rather, it evolves into a continuous process of refinement and adaptation. To ensure consistent and efficient performance, these models necessitate training on vast datasets that encapsulate a diverse spectrum of human activities and environmental conditions. This research endeavor thus assumes the crucial mission of crafting more advanced architectures and techniques to further augment the accuracy and robustness of human activity detection in surveillance systems. By doing so, we

aim to fortify the technological bulwark that safeguards our communities, enriching the field of surveillance with cutting-edge innovations and safeguarding public safety for generations to come.

2.2 Deep Learning

In the realm of machine learning, there were inherent limitations that hindered its ability to handle vast amounts of data. While machine learning excelled with smaller datasets, as data gradually increased, deep learning emerged to address more complex problems and operations.

Regarding Feature Extraction: Unlike machine learning, where manual feature input is required, deep learning autonomously learns and generates high-order features crucial for identification.

What?

Handling extensive volumes of both structured and unstructured data.

Where?

Applications extend to robotics, bioinformatics, and various other fields.

Taking a step back to the 1960s, post-World War II, companies shifted their focus to computer and space research. During this time, a visionary introduced the concept of perceptrons, suggesting a mechanism for learning similar to human neurons. Fast forward to 1969, Marvin Minsky, a researcher, identified a flaw in perceptrons. His published paper highlighted the inability of perceptrons to learn the XOR function, leading to the decline of the perceptron era also known as the AI winter. In the 1980s, Joff Hinton, recognized as the Father of Deep Learning, conducted research on

perceptrons and neurons while working at Google. His landmark paper, "Learning Representations Using Backpropagation Errors," demonstrated that a single perceptron could converge to a linear function but struggled with nonlinear functions. However, a collection of perceptrons could effectively converge to any nonlinear function. After 1991, challenges persisted—neural networks struggled with large datasets due to issues such as a lack of labeled data, limited computational power for additional layers, and difficulties initializing weights. Other machine learning algorithms, like Random Forest, outperformed neural networks on smaller datasets. In 2006, Geoffrey Hinton published the paper "Unsupervised Pre-training of Deep Belief Networks," introducing techniques for weight initialization. This marked the inception of Unsupervised Pre-training, demonstrating that by adding more layers, deep belief networks could be constructed. By 2012, deep learning gained prominence when Joe Felter and his team participated in the ImageNet challenge, achieving significantly reduced error rates by leveraging GPUs for deep learning. Subsequently, major technology companies, collectively referred to as FANG (Facebook, Amazon, Netflix, Google), delved into deep learning research. Google began exploring self-driving cars, while DeepMind, in 2016, embarked on the development of an artificial intelligence game called 'Go player.'

2.3 Types of Neural Networks:

In supervised learning (classification and regression), neural networks excel in capturing nonlinear relationships. Adding hidden layers enhances the network's capability to capture meaningful features.

2.3.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) represent a wondering category of deep neural networks designed explicitly for processing structured grid-like data, such as images or sequences. Their application has profoundly impacted the field of computer vision, becoming a cornerstone for tasks like object detection, image classification and image segmentation. The fundamental concept driving CNNs is the use of convolutional layers, allowing the network to autonomously acquire hierarchical representations of input data.

Here's an overview of how CNNs operate:

2.3.1.1 Convolutional Layers

- CNNs comprise one or more convolutional layers, each applying a set of learnable filters (kernels) to the input data.

- Filters, small-sized matrices, traverse the input data, executing element-wise multiplication and summation operations to generate a feature map.

- The convolutional process captures local patterns and spatial dependencies by leveraging shared weights within the filters.

- The feature maps in resulting spatial dimensions are influenced by the filter size, stride, and padding.

2.3.1.2 Pooling Layers

- Often introduced after convolutional layers, pooling layers reduce feature map dimensions while retaining essential information.

- Pooling divides feature maps into non-overlapping regions and aggregates them, typically via max pooling or average pooling.

- Pooling contributes to achieving translation invariance, resilience to spatial variations, and decreasing computational complexity.

2.3.1.3 Non-linear Activation Functions

- Following each convolutional or pooling layer applied, activation function (e.g., ReLU, LeakyReLU) in case non-linearity

- This non-linearity based activation functions, enable the network to learn relationships and capture high-level features.

2.3.1.4 Fully Connected Layers

- Towards the CNN architecture, one or more fully connected layers are typically incorporated.

- These layers act as classifiers, mapping high-level features learned earlier to the desired output.

2.3.1.5 Training:

- CNNs undergo training through backpropagation, where to minimize loss function the gradients with respect to network parameters guide weight updates via optimization algorithms like stochastic gradient descent (SGD).

- Training occurs on large labeled datasets, enabling the network to automatically extract meaningful and discriminative features from input data through repeated exposure to training examples.

The hierarchical and localized learning capabilities of CNNs make them highly effective for image-related tasks. Convolutional layers, non-linear activation, and pooling layers

allows CNNs to automatically learn features at varying levels of abstraction, from basic edges to intricate object representations. This capability positions CNNs as leading performers in diverse computer vision tasks, establishing their indispensability in the realm of deep learning.

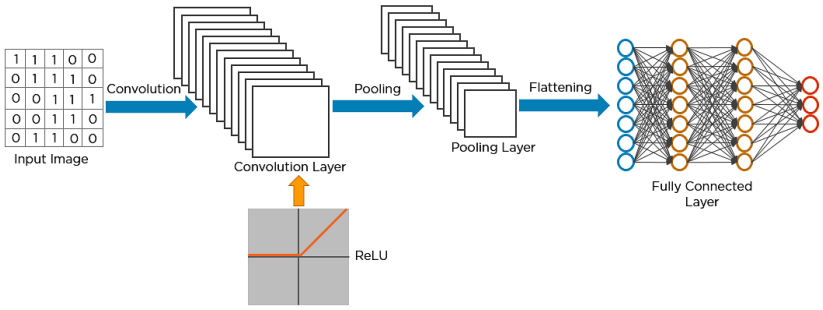

Figure 3: Convolutional Neural Network

2.4 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a type of neural network specifically designed for handling sequential data, such as time series, text, speech, or any data with temporal dependencies. They take input once, and have feedback connections that allow information to share across different time steps. This enables RNNs to model and capture temporal dependencies in the data.

The key concept of RNNs is the recurrent connection, which allows the network to maintain memory that can be updated and influenced by the current input as well as the previous state. This memory enables RNNs to process sequential data by incorporating information from past time steps while considering the current input. Here's how RNNs work:

2.4.1 Recurrent Connections

- In an RNN, each neuron has a recurrent connection that feeds its own output back as an input for the next time step.

- This feedback loop allows the network to maintain a memory or hidden state that summarizes the information seen so far.

- The hidden state captures the context or representation of the past inputs and influences the processing of the current input.

2.4.2 Time Unrolling

- To facilitate understanding and computation, RNNs are often depicted as unrolled over time.

- Each unrolled step represents the RNN processing at a particular time step, with inputs and outputs flowing sequentially.

2.4.3 Training and Backpropagation

- RNNs are trained using backpropagation through time (BPTT).

- BPTT involves unfolding the RNN over time, treating it as a deep feedforward neural network, and applying the standard backpropagation algorithm.

- The gradients are backpropagated through time, allowing the network to learn and update the weights that influence the hidden state and the output predictions.

2.4.4 Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs)

- Traditional RNNs suffer from problem of the vanishing gradient, limiting their ability to capture long-term dependencies in the data. That’s RNNs enable to learn long sequences.

- To address vanishing gradient problem, the Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), comes in picture.

- LSTM and GRU networks include additional mechanisms, known as gates, that selectively control the flow of information, preventing the vanishing or exploding gradients.

- These gated architectures have improved the ability of learning for long-term time, making them more effective in tasks requiring long-range contextual information.

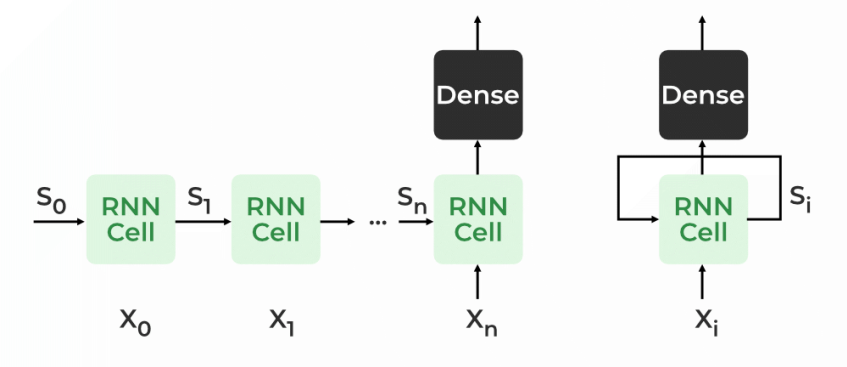

Figure 4: Recurrent Neural Network

RNNs are widely used in different applications, including natural language processing, speech recognition, machine translation, and time series forecasting. These models well-suited for tasks where the context and ordering of the data are crucial. However, it's worth noting that standard RNNs can struggle with very long sequences due to the vanishing gradient problem. In such cases, LSTM or GRU architectures are often preferred.

2.5 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) consists of two components: a generator network and a discriminator network. GANs are designed to generate new data samples and discriminator designed to distinguish fake and real.

Here's how GANs work:

2.5.1 Generator Network

- The generator network takes random input noise (latent space vector) as an input and attempts to generate synthetic data samples.

- It starts with random noise and progressively transforms it into data samples that resemble the training data.

- The network typically comprises several layers, including fully connected or convolutional layers, followed by activation functions.

2.5.2 Discriminator Network

- The discriminator network acts as a binary classifier that distinguishes between real and generated data samples.

- It takes either a real data sample from generated sample from the generator as input and predicts whether it is fake or real.

- The discriminator is trained using real data samples labeled as "real" and generated samples labeled as "fake."

2.5.3 Adversarial Training

- The generator and discriminator networks are trained simultaneously in an adversarial manner.

-The generator's aim is to generate such samples that the discriminator cannot distinguish from real samples.

- The generator and discriminator networks compete against each other, each trying to outperform the other.

2.5.4 Training Process

- The training process alternates between updating the generator and the discriminator networks. During each iteration, a batch of real data samples and a batch of generated samples are used to update the discriminator's weights.

- The generator's weights are updated based on the gradients of the discriminator's decision with respect to the generated samples.

- This adversarial training process continues until the generator produces samples that are cannot distinguish from real samples.

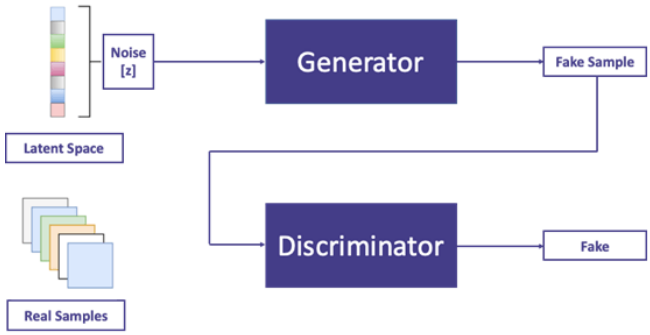

Figure 5: Generative Adversarial Network

The goal of GANs is to train the generator network to learn the underlying distribution of the training data, enabling it to generate new samples that resemble the real data. The generator improves by learning from the feedback provided by the discriminator. As training progresses, the generator gets better at producing realistic samples, while

the discriminator becomes more accurate in distinguishing between real and generated samples.

GANs have gained attention due to their ability to generate highly real and fake data. They have been successfully applied in various domains, including music generation, text generation, and video generation. GANs have also led to advancements in areas such as image-to-image translation, and data augmentation, offering exciting possibilities for creative applications and data generation in deep learning.

CHAPTER 3: LITERATURE REVIEW

The relevant research was conducted on the recognition of suspicious activity [2] where they present overall progress, issues, and challenges. The researchers presented a literature review of 6 abnormal activities and discussed the steps involved in recognizing human activity, including foreground object extraction, feature extraction, and object detection by using non-tracking or tracking methods. The study [3] utilized the KTH and CAVIAR datasets, where features were computed from video frames, and the classifier made predictions. In the study [4] lighting and nighttime monitoring are crucial components and a CNN model using infrared (IR) images achieved 90% accuracy in object recognition. In the context of education [5], detecting cheating activities in examination halls is critical. The SURF (Speeded Up Robust Features) algorithm was used to match and find corresponding features, and Viola-Jones object detectors were employed to detect faces and label activities. Focused on a group of people in a Warfield [6], implemented a real-time algorithm using a motion-based moving object detection method. Another study [7] on violence detection employed a CNN-LSTM model having 89% accuracy by collecting datasets from Google. The system takes the input of real-time videos from CCTV cameras and then passes them to a CNN model which is created by the transfer learning that detects objects and generates alerts. A study [8] in examination halls identified various features such as head, hands, and posture. By identifying these features, the system can detect whether a student is engaged in cheating. Furthermore, the detection of firearms plays a crucial role in crime prevention. deep learning neural network model is introduced [9], focusing on the presence of weapons and individuals wearing helmets while carrying weapons. The system provides enhanced security while

keeping costs minimal. In another research [10] the YOLO-V3 model was utilized for human activity detection, achieving an accuracy of 86.2%. In research [11], the Cov2d-LSTM and ConvLSTM2D models on the UCF Crime Dataset, achieved accuracy rates of 0.68 and 0.64, respectively. An alternative approach for the detection of shoplifting incidents was introduced in [12]. They used 2D CNN, 3D CNN, and a fusion of Inception V3 with BILSTM models and got accuracy 49%, 57%, and 81%, respectively. A multi-stage process, encompassing image preprocessing [13] proposed model is to be effectively trained on training data, thus ensuring robust performance on previously unseen datasets. Furthermore, a Hybrid Machine Learning Algorithm was applied to the [14] UCF crime dataset. They combined the potential of LSTM and CNN for feature extraction. To enhance the model performance, temporal features were further extracted using a multilayer LSTM architecture, getting an impressive accuracy of 96%. An automated deep-learning architecture was introduced [15] which is centered around 3D ConvNets to extract spatiotemporal features. The experimental results showcased an accuracy rate of 82%. In another contribution [16] proposed an approach that achieves the accuracy of 3D-CNN for feature extraction, coupled with the BILSTM model for the classification. In research [17], CNN methodology achieves accuracy 90.2% was employed on a dataset of 2000 images such as blood, guns, and knife. Deep Neural Network was conducted [18], achieving an accuracy of 98%, on crowd violence, UCSD, and violent flow datasets. A combination of CNN and BER model was implemented [19] resulting in an accuracy of 85.63% 2023 International Conference on Open-Source Systems and Technologies (ICOSST) by using the XD-Violence audio and video datasets. Live CCTV data (CAVIAR dataset) is used [20] to employ a CNN-GRU model for the detection of faces and weapons, achieving an

impressive accuracy of 95.97%. In another research [21], Twitter data was subjected to keyword filtering and labeling. The text mining methodology successfully categorized 10 distinct crime classes. By employing SVM and ANN on the dataset the researchers achieved an accuracy of 90.3%. Another study [22] found on the Saudi Arabia tweets (2017-2021), where they identify the keywords that are used for criminal activities. they applied both ML and DL algorithms where they achieved an accuracy of 79%. Chicago crime data used [23] to detect and map a crime-dense region in NYC. The crime prediction research [24] highlights unemployment and literacy rates. They employed a random forest regressor, yielding an accuracy of 97%. In recent advancements, ML and DL models were used to forecast various crime types [25], an impressive accuracy of 99% was achieved when applied to a weather dataset. This paper [31] focuses on implementing smart surveillance systems to detect abnormal behaviors that pose security risks. The study specifically targets two human activities, namely walking and running, without any restrictions on the number of people involved. Various techniques such as background subtraction and segmentation are utilized for this purpose. In the study [32], Deep learning algorithms have been validated to be highly effective in computer vision tasks, achieving superior accuracy. Particularly CNNs offer significant assistance in addressing classification, object detection, and segmentation issues. When classifying human activities as abnormal or normal, convolutional and recurrent neural networks consistently yield better results compared to traditional machine learning methods. In the study [33] Ensemble learning techniques were employed to address crowd-related issues. The implementation resulted in the development of an alert system that effectively records, detects, and reports suspicious activities. This system provides enhanced

security while keeping costs minimal. Human activity detection is a valuable capability across various domains, including autonomous vehicles, where the detection of human gestures plays a significant role. our study has unveiled an impressive level of performance that surpasses that of the recently proposed method [34]. The methodology detailed in reference [35] harnessed the power of Convolutional Neural Networks (CNNs), resulting in Area Under the Curve (AUC) scores ranging from 73% to 77%. Additionally, their combined approach involving CNN with Long Short-Term Memory (CNN-LSTM) yielded AUC scores ranging between 70% and 72%, while the utilization of Stacked LSTM led to AUC scores spanning from 65% to 72%. The research [36] uses VGG-16 model to monitor student behavior in examination where they first extract frames from videos and then apply model to predict student activity. The disadvantage is that the scope of this study is limited only for students so it can be extended because in examination hall other than student activity there may be happens some other crime activities. The research conducted in 2014 where CNN model trained on 1 million dataset which is categories into 487 groups. They also use the UCF-101 dataset where they increase performance from 43.9% to 63.3%. They study conducted by Sathyajit et al. [38] to detect guns in real time. They used captured images for model training which is beneficial in efficiently detecting abandoned baggage. Li et al. [39] conduct study based on geographical location data to detect human activity such as lying, running down, and walking. They trained CNN model and got accuracy 86.7 percent. The 3D CNN used in research [40] to detect suspicious and normal activities both. There are some drawbacks of their research they trained model for binary classification suspicious or normal and then they trained it for multi-

classification. By this approach mode allows false positives between criminal categories.

The relevant research [41] uses data from security cameras and built CNN Alex Net model to detect fall detection. The study by Om et al. [42] trained CNN model to detect Break-In, Shoplifting and Robbery. Their model takes video and then model will extract frames from video. After extracting frames, the trained model will be passed and the predicted labels will be pushed in queue. Thus, based on highest probability model will the system will give answer. The research [43] examines 13 anomalies using the UCF crime dataset. They used the pre-trained model CNN and then process it further using model Bi-directional long short-term memory (BD-LSTM). The semantic based approach discussed in [44], they classified between suspicious and normal but their dataset is not standard public dataset. The [45] discussed spatial and temporal network using CNN. They use spatiotemporal data and gets state-of-the-art results. Another research [46] investigate the semantic based recognition system where object detection, object tracking and classification was conducted. Their main goal is to detect human motion.

The study [47] worked on object tracking using raw video data. The system classifies the labels after tracking objects in 2D and lastly 3D features were calculated. Another study [48], proposed Single Shot Detector (SSD) algorithm and for training they use usual and unusual activities. The wide study is available [49], which worked on detection, tracking and description of behaviors, personal identification. They face issues in detecting human habits. The research [50] based on inter-object motion features. Their system detects various type of behaviors.

In the proposed work, we detect three suspicious activities (Fighting, Shoplifting and Robbery). A lot of studies conducted to detect activities for video surveillance but none of have accurate accuracies.

CHAPTER 4: METHODOLOGY AND FRAMEWORK

3.1 DATA ANALYSIS

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

3.1.1 Data Pre-Processing

The data pre-processing steps carried out on the UCF Crime Dataset (see figure3), which is crucial in preparing the data for further analysis, particularly in the context of deep learning and LSTM models.

Figure 6: UCF Crime Dataset

Dataset Selection: First, we choose the UCF Crime Dataset, focusing on three common crime categories; Robbery, Shoplifting, and Fighting. These categories will likely be the classes that your deep learning model will aim to classify images into.

Data Splitting: The dataset is further divided into Train and Test subsets. This common practice allow to assess how well the model generalizes to new, unseen data.

Image Characteristics: The images in the dataset have dimensions of 64x64 pixels and are stored in (.png) format. You mention that they were extracted from video footage using a systematic sampling approach, where every 10th frame was chosen from each video.

Gray scale Conversion: As a data preprocessing step, the images are converted to grayscale. Grayscale images have only one channel and are less computationally intensive compared to full-color images. This conversion retains intensity information while reducing complexity.

Resizing: To ensure efficient batch processing during model training, you resize all the images to a consistent size of 50x50 pixels. Uniform image sizes are often necessary for deep learning models.

Overall, your data pre-processing pipeline is comprehensive and tailored to the specific requirements of your deep learning model, which combines CNN and LSTM architectures for image recognition and activity detection in surveillance footage. These steps help ensure that the data is in the right format for training and testing your model effectively.

3.1.2 Feature Selection

In the context of our experiments, feature selection was a crucial consideration to enhance the effectiveness of our deep learning models. While deep learning models are adept at feature learning from raw data, we recognized the value of optimizing the input features for our specific problem. We conducted feature selection to identify and retain the most informative attributes that would contribute to the model's performance.

This process aimed to reduce dimensionality, eliminate noise, and improve model interpretability. By carefully selecting and engineering features, we were able to create a more streamlined and efficient input data representation for our CNN and LSTM models. This approach ensured that the models were focused on the most

relevant information, ultimately leading to improved accuracy and robustness in our experiments.

In our study, we suggest using a tool called Oriented FAST and Rotated BRIEF (ORB) to work with images. ORB is like a detective for pictures. It spots special points in an image and figures out what's unique about them. This helps us describe the important features of an image in a simple way. While other methods like SURF, SIFT, and HOG have been used in similar studies, we picked ORB because it's quicker and can find about 300 special points in an image **[52]**.

In the data preparation phase, the first step involves loading and processing image data using the OpenCV library and the ORB feature descriptor. The 'load-data' function is employed to load image data for different categories, with each category associated with a label. Subsequently, all images are resized to a consistent size of 50x50 pixels to facilitate efficient batch processing during model training. The ORB detector is then applied to extract features, wherein the script sets the score type to ORB-HARRIS-SCORE (which can be experimented with). The `detectAndCompute` method is used to obtain keypoints and descriptors for the resized image. Keypoints represent distinctive points in the image, while descriptors capture information around these keypoints.

Once our detective has found these points, we draw them on the image so we can see where they are. We call this the 'image-with-keypoints'. The special part is that these special points and their descriptions help us understand the images better. It's like having clues that make it easier for computers to recognize and understand pictures. This can be useful for different things like making sure two images are similar, figuring out what's in a picture, or even tracking objects in videos.

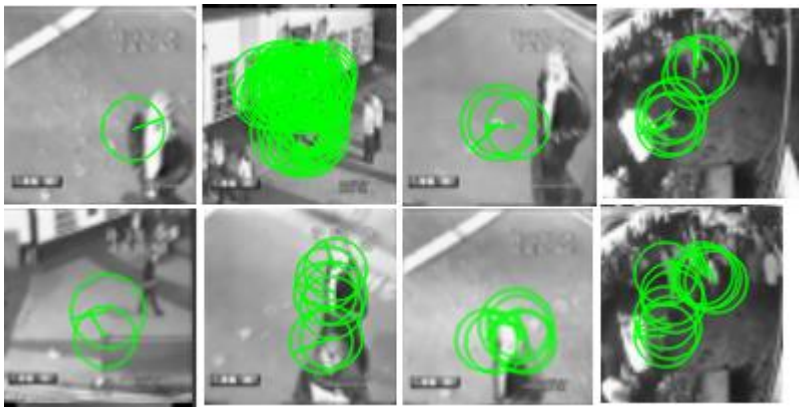

Figure 7-Visualization with SURF (Speed-up Robust Features)

3.2 Model Selection

In our experiments, we employed two key deep learning components, the Long Short-Term Memory (LSTM) model and the Convolutional Neural Network (CNN), to tackle the challenge of detecting and classifying criminal activities in surveillance footage.

3.2.1 Long Short-Term Memory (LSTM)

LSTM model, constructed using Keras-based LSTM layers, played a pivotal role in understanding temporal dependencies within the input data. The model consisted of two LSTM layers, each comprising 8 units. The first LSTM layer was particularly adept at retaining sequences, allowing it to capture temporal information from a substantial 2500-time steps, with each step characterized by a single feature and utilizing the hyperbolic tangent (tanh) activation function. Subsequently, a dense layer with 4 neurons further refined the model's ability to learn complex relationships within the data. To enhance generalization, we introduced a 20% dropout layer. Finally, a flattened layer was employed to reshape the 3D output from the preceding LSTM layers into a 1D vector, preparing the data for the classification task. The LSTM model's strength in processing sequential data and capturing temporal patterns seamlessly

complemented the CNN component in our comprehensive approach to detecting and classifying criminal activities. The integration of both the CNN and LSTM elements resulted in exceptional accuracy, significantly contributing to the enhancement of security measures and public safety.

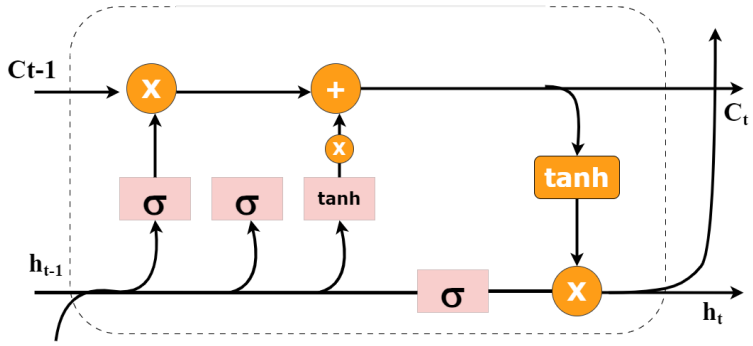

Figure 8: Long-Short Term Memory

3.2.2 Convolution Neural Network (CNN)

Convolution Neural Network (CNN) This model employed in our research for activity detection is tailored to efficiently process grayscale images with dimensions of 50x50 pixels. The network is combination of multiple layers, each serving a specific purpose in the extraction of essential features and subsequent classification. The process commences with three successive convolutional layers, equipped with 64, 128, and 256 filters, respectively, each with a filter size of 3x3 pixels. These convolutional layers effectively capture crucial patterns and distinctive characteristics from the input images, generating output feature maps with varying spatial dimensions. To introduce non-linearity and enhance the model's learning capacity, Leaky ReLU activation functions follow each convolutional layer. Below you can architecture of the CNN model. Subsequent to the convolutional layers, max-pooling layers are strategically inserted, operating with a pool size of 2x2 pixels. These layers facilitate the down

sampling of feature maps, reducing their spatial dimensions while retaining essential information. To mitigate overfitting the model generalization capabilities, dropout layers are thoughtfully incorporated after each max pooling layer. During training, these dropout layers randomly deactivate 25% and 40 Following the last convolutional layer, the output feature maps the flattened into a 1D vector, preparing them for further processing through fully connected layers. A singular fully connected layer with 256 neurons is introduced, followed by a Leaky ReLU activation function and a dropout layer with a dropout rate of 50%. This fine-tunes the extracted features and introduces additional regularization to the model. The final layer of the CNN model, known as the output layer, comprises three neurons, representing the three distinct classes: Fighting, Shoplifting, and Robbery. A softmax activation function is applied in the output layer, converting the raw scores into probabilities, thus facilitating the classification of criminal activities based on the highest probability. Overall, the CNN architecture adeptly captures relevant features from the grayscale images, effectively discerning distinctive patterns characterizing various criminal activities. The use of activation functions and dropout layers having the ability of effective regularization, ensures that will work well on unseen data. This well-structured architecture, when combined with our LSTM model, showcases remarkable accuracy in identifying and classifying criminal activities in surveillance footage, thereby significantly enhancing security and safety measures.

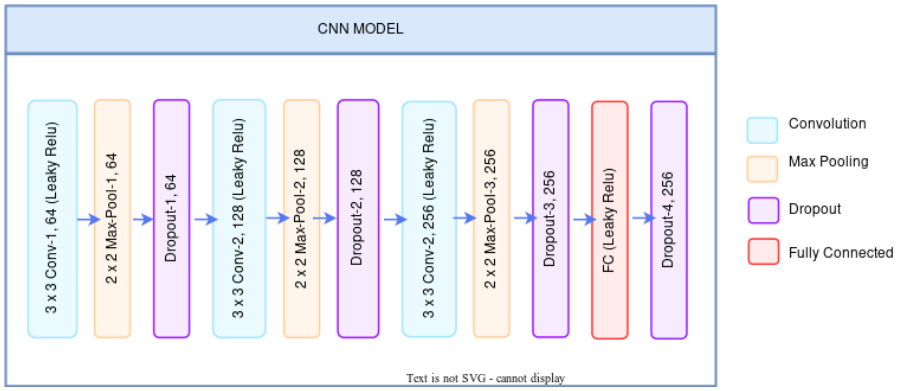

Figure 9: Convolutional Neural Network

3.2.3 Combine Architecture CNN-LSTM

we combine a CNN model and an LSTM model to create a hybrid architecture (See Figure 7) for a classification task involving three classes. We have achieved a significant milestone by seamlessly integrating the CNN and LSTM models. This ingenious fusion allows the model to simultaneously process image data and sequential data, capitalizing on CNN expertise in extracting visual features from grayscale images and the LSTM proficiency in capturing temporal patterns in sequences.

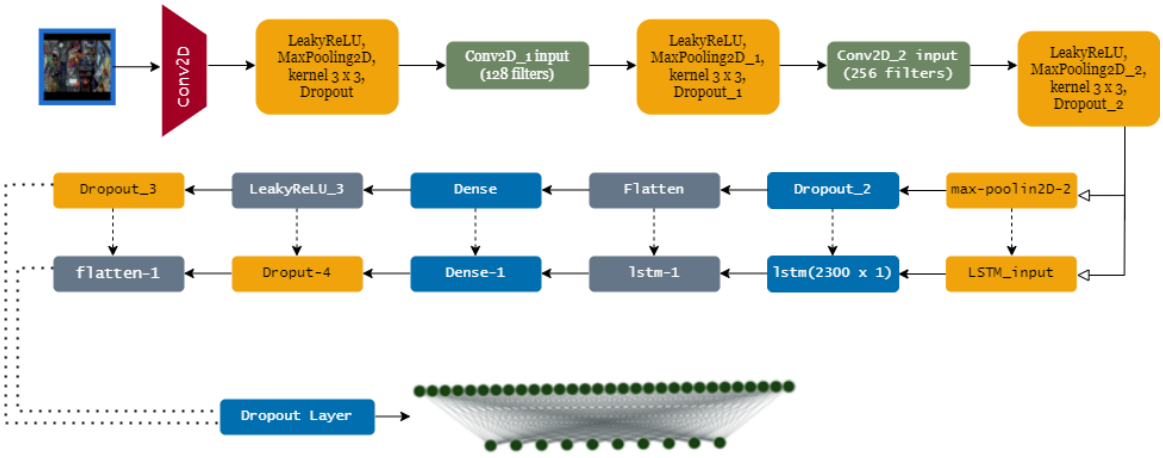

Figure 10: Mono-Scale-CNN-LSTM Architecture

For the assembly of the model, we adopt categorical cross-entropy as the designated

$$\boldsymbol{Loss} = \sum \left(\mathrm{y_i} \times \log_p i\right) \quad \textbf{(1)}$$

where $\mathrm{y_i}$ signifies the genuine class label and pi denotes the predicted probability for class i. To drive optimization, the Adam optimizer is harnessed, characterized by its dynamic learning rate.

$$\boldsymbol{Optimization} = Adam\,(learning\ rate = 0.001, \beta 1 = 0.9, \beta 2 = 0.999, \epsilon = 1e - 07) \quad \textbf{(2)}$$

To monitor the model's training progress effectively, two callbacks are employed, to ensure supervision during the learning process. Throughout the training process, both CNN and LSTM inputs are provided to the model. With a batch size of 64 and twenty epochs, we synergize the strengths of both networks to optimize activity detection performance. The CNN model processes the input images, and the LSTM model handles sequential data, with its cell computations given by formulas described earlier. Validation is performed using test data to assess the model's generalization ability. Upon the conclusion of training, we rigorously evaluate the hybrid model's performance through the evaluate method, which computes the validation loss and accuracy on the test data. An impressive accuracy of 99% signifies the model's exceptional capability in accurately identifying and classifying criminal activities in surveillance footage, bolstering security measures, and enhancing public safety significantly. The integration of CNN and LSTM models within the hybrid architecture marks a significant breakthrough in activity detection systems. With its remarkable accuracy and robustness, this model exhibits tremendous potential for real-world surveillance applications, equipping security personnel with a formidable tool to maintain safer and more secure environments.

CHAPTER 5: RESULTS

4.1 Model Evaluation and Comparison

Research closely monitored the learning progress of the integrated CNN-LSTM model across multiple epochs. By using Oriented FAST and Rotated BRIEF (ORB) the training accuracy steadily increased, reaching an impressive value of 0.99. An increase in accuracy indicates that the model achieved high accuracy in correctly classifying each category samples from the training dataset. There are numerous metrics commonly employed to assess the performance of classifiers. This research measures the performance using accuracy, precision, recall and f1-score.

Accuracy gauges the ability of a trained classifier to correctly predict class labels in comparison to the actual labels.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(3)}$$

$$Precesion = \frac{TP}{TP+FP} \quad \textbf{(4)}$$

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(5)}$$

$$F1 - score = \frac{2 \times Precesion \times Recall}{Precesion+Recall} \quad \textbf{(6)}$$

The precision, recall, F1-score, and accuracy of algorithms for the crime dataset are shown in Table 1. Similarly, the validation accuracy also showed improvement, reaching a value of 1.00. The validation accuracy (See Figure 11) is computed based on the model's performance on a separate validation dataset, and an accuracy of 99% suggests that the model generalizes well to unseen data and can accurately classify samples from the validation dataset.

Table 1: Architecture details about the proposed model

Block	Input	Kernel	Output
Conv2D_input	50 x 50 x 1	3 x 3 x 3	50 x 50 x 1
conv2d	50 x 50 x 1	---	50 x 50 x 64
LeakyReLU	50 x 50 x 64	---	50 x 50 x 64
MaxPooling2D	50 x 50 x 64	2 x 2 x 2	25 x 25 x 64
Dropout	25 x 25 x 64	---	25 x 25 x 64
conv2d_1	25 x 25 x 64	3 x 3 x 3	25 x 25 x 128
LeakyReLU	25 x 25 x 128	---	25 x 25 x 128
MaxPooling2d_1	25 x 25 x 128	2 x 2 x 2	13 x 13 x 128
dropout_1	13 x 13 x 128	---	13 x 13 x 128
conv2d_2	13 x 13 x 128	3 x 3 x 3	13 x 13 x 256
leaky_re_lu_2	13 x 13 x 256	---	13 x 13 x 256
Max_pooling2d_2, lstm_input	13 x 13 x 256, 2500 x 1	2 x 2 x 2	7 x 7 x 256, 2500 x 1
dropout_2, lstm	7 x 7 x 256, 2500 x 1		7 x 7 x 256, 2500 x 8
Flatten, lstm_1	7 x 7 x 256, 2500 x 1		None x 12544, 2500 x 8
Dense, dense_1	None x 12544, 2500 x 8		None x 256, 2500 x 4

leaky_re_lu_3, dropout_4	None x 256, 2500 x 4	None x 256, 2500 x 4
droput_3, flatten_1	None x 256, 2500 x 4	None x 256, 100000
concatenate	256, 100000	10256
dense_2	10256	3

Throughout the training process, the model's performance was optimized by minimizing the training loss. The training loss steadily decreased during training and reached a minimum value of 5. The training loss is calculated using a loss function, such as categorical cross-entropy. Similarly, the validation loss (Figure 12), which evaluates the model's performance on the validation dataset, also decreased during training and reached a minimum value of 10. A lower validation loss indicates that the model generalizes well and makes accurate predictions on unseen data.

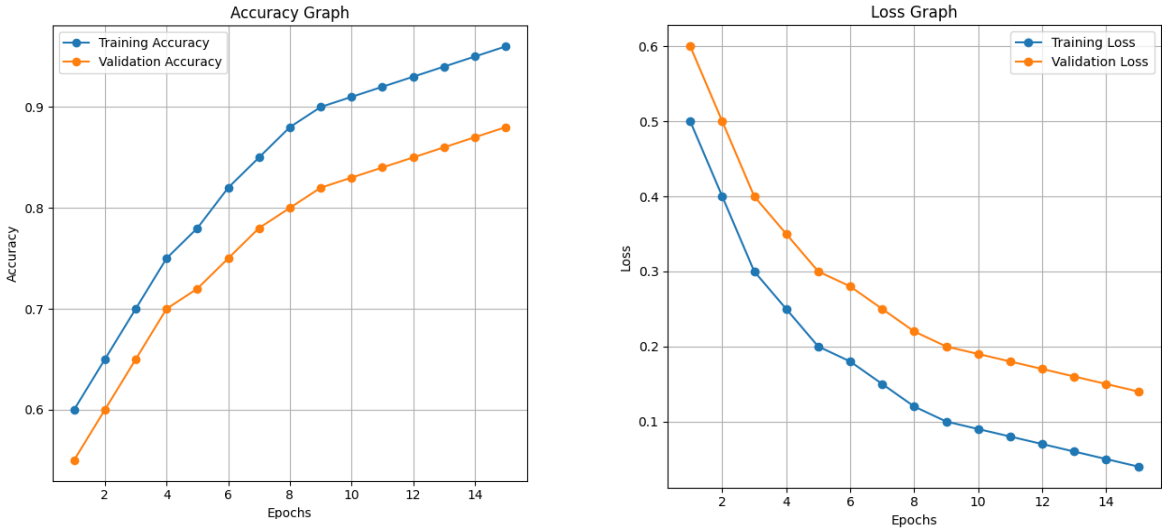

Figure 12: Loss of Proposed Architecture CNN-LSTM on UCF Dataset

Figure 11: Accuracy of Proposed Architecture CNN-LSTM ON UCF Dataset

In the confusion matrix before normalization (Figure 13), model prediction for each class ('Fighting', 'Shoplifting', and 'Robbery') was analyzed. After normalizing the confusion matrix (Figure 14), the counts are transformed into percentages, providing a more comprehensive view of the model's performance, considering the varying sizes of the classes.

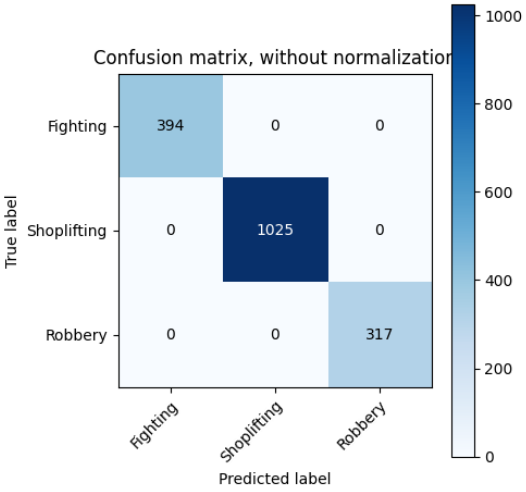

Figure 13: Confusion Matrix without Normalization

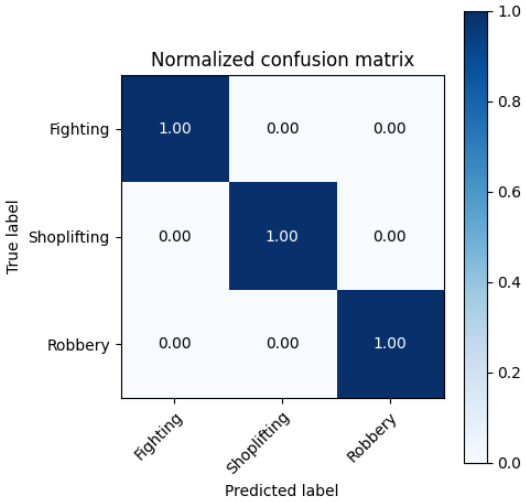

Figure 14: Confusion Matrix with Normalization

The normalized confusion matrix revealed the following rates of correct classifications for each class: 100% for 'Fighting', 100% for 'Shoplifting', and 100% for 'Robbery'. These percentages indicate the model's proficiency in accurately classifying instances within each category. The higher percentages on the diagonal of the matrix indicate strong performance in accurately identifying each activity. However, off-diagonal elements represent misclassifications. Despite this, the model achieved commendable accuracy, with a 94% correct classification rate for 'Shoplifting' and 87% for 'Robbery'. Overall, the high training and validation accuracy, decreasing training and validation losses, and the normalized confusion matrix provide valuable insights into the model's performance. These results demonstrate the model's ability to accurately detect and classify criminal activities in surveillance footage, highlighting its potential for real-world applications in enhancing security measures and ensuring public safety.

Through the comparison of the presented model with alternative approaches, we achieved an impressive accuracy rate of 99%, which notably surpasses previous

results. The analysis is elaborated in Table 1, where diverse algorithms such as DenseNet121, Custom CNN, CNN-LSTM, and VGG16, etc., are evaluated.

Table 2: Applied Models Accuracy Results on Allahabad-2022 Dataset

Model	Precision	Recall	f1- score	Accuracy
Custom CNN	0.81	0.59	0.62	77%
DenseNet121	0.83	0.78	0.80	88%
VGG19	0.81	0.76	0.77	76%
DenseNet201	0.70	0.79	0.74	73%
ResNet50	0.34	0.59	0.43	59%
ResNet101V2	0.70	0.50	0.55	50%
ResNet50V2	0.70	0.50	0.55	50%
ResNet152	0.60	0.40	0.55	45%
DenseNet169	0.88	0.83	0.84	83%
ResNet152V2	0.64	0.25	0.13	25%
ResNet101	0.33	0.18	0.05	18%
CNN-LSTM	**1.00**	**0.95**	**0.98**	**98%**

Table 3: Applied Models Accuracy Results on Pre-Crime Dataset

Model	Precision	Recall	f1- score	Accuracy
Custom CNN	0.81	0.59	0.62	77%
DenseNet121	1.00	1.00	1.00	99%
VGG19	0.26	0.51	0.35	51%
DenseNet201	0.70	0.79	0.74	73%
ResNet50	0.96	1.00	0.98	97%
ResNet101V2	1.00	1.00	1.00	100%
ResNet50V2	0.85	0.96	0.90	90%
ResNet152	0.26	0.51	0.35	51%
DenseNet169	0.50	1.00	0.66	50%
ResNet152V2	0.99	0.99	0.99	99%
ResNet101	0.71	0.66	0.63	66%
CNN-LSTM	**1.00**	**1.00**	**1.00**	**99%**

Table 4: Applied Models Accuracy Results on UCF Crime Dataset

Model	Precision	Recall	f1- score	Accuracy
Custom CNN	0.81	0.59	0.62	77%
DenseNet121	0.90	0.86	0.87	86%
VGG16	0.84	0.81	0.76	81%
VGG19	0.81	0.76	0.77	76%
DenseNet201	0.77	0.71	0.71	71%
ResNet50	0.34	0.59	0.43	59%
ResNet101V2	0.70	0.50	0.55	50%
ResNet50V2	0.70	0.50	0.55	50%
ResNet152	0.60	0.40	0.55	45%
DenseNet169	0.23	0.27	0.16	27%
ResNet152V2	0.64	0.25	0.13	25%
ResNet101	0.33	0.18	0.05	18%
CNN-LSTM	**1.00**	**1.00**	**1.00**	**99%**

CHAPTER 5: CONCLUSION

Our primary objective is to enhance the accuracy of the presented model using feature extraction technique Oriented FAST and Rotated BRIEF (ORB) while also reducing its susceptibility to variations. This research introduced a CNN-LSTM approach that achieves high accuracy in abnormal activity detection. The research approach involves processing image frames, extracting features, and then recognizing them using a

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

REFERENCES

[1] “https://ocindex.net/explorer.”

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

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

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

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

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

[7] Gurmeet Kaur and Sarbjeet Sing, “Advances in information communication technology and computing,” 2022, pp. 1–6.

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

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

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

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

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

[13] Muneer, I., Saddique, M., Habib, Z., Mohamed, and H. G., “Shoplifting detection using hybrid neural network cnn-bilsmt and development of benchmark dataset,” in Applied Sciences, 2023.

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

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

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

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

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

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

[20] R. Shenoy, D.Yadav, and H.Lakhotiya and J.Sisodia, “An intelligent framework for crime prediction using behavioural tracking and motion analysis,” 2022, pp. 1–6. [21] M. A. Permana, M. I. Thohir, and T. Mantoro and M. A. Ayu, “Crime rate detection based on text mining on social media using logistic regression algorithm,” 2021, pp. 1–6.

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

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

[24] L. G. A. Alves and H. V. Ribeiro and F. A. Rodrigues, “Crime prediction through urban metrics and statistical learning,” vol. 505, 2018, pp. 435–443. [25] V. M. L. Elluri and N. Roy, “Developing machine learning based predictive models for smart policing,” 2019, pp. 198–204.

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

[26] Beddiar, D.R., Nini, B., Sabokrou, M. and Hadid, A.; Vision-based human activity recognition: a survey. Multimedia Tools and Applications, 79(41), pp.30509-30555, 2020.

[27] Varshney, P., Harsh Tyagi, N.K., Lohia, A.K. and Girdhar, P.; A Deep Learning Based Approach to Detect Suspicious Weapons. Proceedings http://ceur-ws. org ISSN, 1613, p.0073, 2021.

[28] Munagekar, M.S.; Smart Surveillance system for theft detection using image processing. International Research Journal of Engineering and Technology (IRJET), 5(8), pp.232-234, 2018.

[29] Arutha, C.V., Jyotsna, C. and Amudha, J.; March. Deep learning approach for suspicious activity detection from surveillance video. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 335-339). IEEE, 2020.

[30] Kakadiya, R., Lemos, R., Mangalan, S., Pillai, M. and Nikam, S.; Ai based automatic robbery/theft detection using smart surveillance in banks. In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 201-204). IEEE, June-2019.

[31] Baile, P., Sutar, N., Shinde, S., Brahmankar, A., Angadi, S., Dhore, P. (2022). A Survey on Suspicious Activity Detection in Examination Hall. Journal of Pharmaceutical Negative Results, 7565-7571.

[32] F. G. Ibrahim Salem, R. Hassanpour, A. A. Ahmed and A. Douma, "Detection of Suspicious Activities of Human from Surveillance Videos," 2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA, Tripoli, Libya, 2021, pp. 794-801, doi: 10.1109/MISTA52233.2021.9464477.

[33] A. M. Bhugul and V. S. Gulhane, "Novel Deep Neural Network for Suspicious Activity Detection and Classification," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-7, doi: 10.1109/SCEECS57921.2023.10063130.

[34] K. Geng and G. Yin, "Using Deep Learning in Infrared Images to Enable Human Gesture Recognition for Autonomous Vehicles," in IEEE Access, vol. 8, pp. 88227-88240, 2020, doi: 10.1109/ACCESS.2020.2990636.

[35] Y. Heryadi and H. L. H. S. Warnars, "Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM," 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Phuket, Thailand, 2017, pp. 84-89, doi: 10.1109/CYBERNETICSCOM.2017.8311689.

[36] Amrutha C.V, C. Jyotsna, Amudha J. : “Deep Learning Approach for Suspicious Activity Detection from Surveillance Video” Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) IEEE Xplore Part Number: CFP20K58- ART; p:335-339.

[37] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014). Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732.

[38] Sathyajit Loganathan, Gayashan Kariyawasam, Prasanna Sumathipala : “Suspicious Activity Detection in Surveillance Footage ” 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA).

[39] Li J, Wu R, Zhao J, Ma Y (2017) Convolutional neural networks (CNN) for indoor human activity recognition using Ubisense system. In: 2017 29th Chinese control and decision conference (CCDC), pp 2068–2072. IEEE.

[40] Guillermo Arturo Martınez, Mascorro, Jose ,CarlosOrtiz-Bayliss, Hugo Terashima-Marın : “Detecting Suspicious Behavior on Surveillance Videos: Dealing with Visual Behavior Similarity between Bystanders and Offenders” University of Gothenburg.

[41] Anishchenko, L (2018) Machine learning in video surveillance for fall detection. In: 2018a ural symposium on biomedical engineering, radioelectronics and information technology (usbereit), pp 99–102. IEEE.

[42] Om M. Rajpurkar, Siddesh S. Kamble, Jayram P. Nandagiri and Anant V. Nimkar “Alert generation on detection of suspicious activity using transfer learning” 11th ICCCNT - IIT – Kharagpur.

[43] Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW (2021) CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimed Tools Appl 80:16979–16995

[44] Guillermo Arturo Martınez, Mascorro, Jose ,CarlosOrtiz-Bayliss, Hugo Terashima-Marin:" IEEE - Detecting Suspicious Behavior on Surveillance Videos: Dealing with Visual Behavior Similarity between Bystanders and Offenders" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[45] Feichtenhofer C, Pinz A, Zisserman A (2016a) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941.

[46] Sandesh patil,kiran talele:" Suspicious movement detection and tracking based on color histogram" 2015 International Conference on Communication, Information & Computing Technology (ICCICT) (15-17).

[47] Mohannad Elhamod,Martin D. Venine:" Automated Real-Time Detection of Potentially Suspicious Behavior in Public Transport Areas" IEEE Transactions on Intelligent Transportation Systems .p:688-699.

[48] Ajeet Sunil , Manav Hiran Seth , Shreyas E:" Usual and Unusual Human Activity Recognition in Video using Deep Learning and Artificial Intelligence for Security Applications" 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT).

[49] Weiming Hu, Tieniu Tan,Liang Wang:" A survey on visual surveillance of object motion and behaviors" IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).p:334-352.

[50] Mohannad Elhamod,Martin D. Venine:" Real-Time Semantics-Based Detection of Suspicious Activities in Public Spaces"2012 Ninth Conference on Computer and Robot Vision.p:268-275.

[51] Sowmeya, V., & Karthik, R. J. Internet of things module accelerated dense deep learning for crime detection in surveillance systems. *International Journal of Health Sciences*, (I), 6364-6379.

[52] K. K. E. Rublee, V. Rabaud and G. Bradski, ""orb: An efficient

alternative to sift or