P.E.S COLLEGE OF ENGINEERING,

MANDYA

(An Autonomous Institute under Visvesvaraya Technological University, Belagavi)

A DISSERTATION REPORT ON

"HUMAN ACTIVITY DETECTION FOR SURVEILLANCE"

Submitted in partial fulfilment of the requirement
For the award of the
BACHELOR OF ENGINEERING DEGREE

Submitted by

Madan Arya M A	**4PS20EC061**
Manish H D	**4PS20EC065**
Manoj M A	**4PS20EC066**
Rakesh M A	**4PS20EC098**

Under the guidance of

Mr. M Subramanyam
Associate Professor

Department of Electronics and Communication Engineering
P.E.S. College of Engineering, Mandya.

2023-2024

P.E.S COLLEGE OF ENGINEERING

MANDYA-571401

(An Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that,

Madan Arya M A	**4PS20EC061**
Manish H D	**4PS20EC065**
Manoj M A	**4PS20EC066**
Rakesh M A	**4PS20EC098**

have successfully completed the project work entitled **"Human Activity Detection for Surveillance"** in partial fulfillment for the award of degree of **Bachelor of Engineering in Electronics and communication Engineering** of **P.E.S college of Engineering, Mandya, VTU Belagavi** during the year **2023-2024**. It is certified that all corrections/suggestions indicated in internal assessment have been incorporated in the report deposited in the Library. The project has been approved as it satisfies the academic requirements in respect of project work prescribed for the degree **in Bachelor of Engineering.**

Signature of the guide
M Subramanyam
Associate Professor

Signature of the HoD
Dr. Punith Kumar M B
Professor & HOD

Dr. H M Nanjundaswamy
Principal
PES College of engineering, Mandya

Project work viva-voice examination			
Sl.No	Examiners		Date
	Name	Signature	
1			
2			

P.E.S COLLEGE OF ENGINEERING

(An Autonomous Institute under VTU, Belagavi)

MANDYA – 571401

Department of Electronics and Communication Engineering

DECLARATION

We **MADAN ARYA M A, MANISH H D, MANOJ M A, RAKESH M A** students of 8th semester Bachelor of Engineering in Electronics and Communication Engineering , PESCE, Mandya, hereby declare that the project work being presented in the dissertation entitled **"Human Activity Detection for Surveillance"** is an authentic record of the work that has been independently carried out by us and submitted in partial fulfillment of the requirements for the award of degree in **Bachelor of Engineering in Electronics and Communication Engineering,** affiliated to **Visvesvaraya Technological University (VTU), Belagavi** during the year 2023-2024.

The work contained in the thesis has not been submitted in part or full to any other university or institution or professional body for the award of any other degree or any fellowship.

Place: Mandya

Date: 29/05/2024

MADAN ARYA M A

MANISH H D

MANOJ M A

RAKESH M A

ACKNOWLEDGEMENT

We feel greatly to express our humble feelings of thanks to one and all that have helped me directly or indirectly in the successful completion of the project work.

We are grateful to our guide **M. Subramanyam,** associate professor**,** Dept of ECE, PESCE, Mandya for providing me a congenial environment to work in and helped me a lot in making this project report, guidance and moral support throughout the project work.

We would like to thank **Dr. Punith Kumar M B,** Professor and HoD, Dept of ECE, and **Dr. H M Nanjundaswamy,** Principal**,** PESCE, Mandya for providing me a congenial environment to work in and helped me a lot in carrying out this project work, guidance and moral support throughout the work.

Finally, we would like to thank all the professors of Dept of ECE, all my friends, who with their constant and creative criticism, made us to maintain standards throughout my endeavor to complete this project work.

We are grateful to our institution **PES College Of Engineering** and **Department Of Electronics And Communication Engineering for** imparting me the knowledge with which we can do our best.

Madan Arya M A 4PS20EC061
Manish H D 4PS20EC065
Manoj M A 4PS20EC066
Rakesh M A 4PS20EC098

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS

HAR: Human Activity Recognition

CNN: Convolution Neural Network

LSTM: Long Short-Term Memory

RNN: Recurrent Neural Networks

UCF50: University of Central Florida 50

BVSB: Best-versus-Second-Best

UAH: Unmanned Aerial Vehicles

PSO: Particle Swarm Optimization

ABC: Artificial Bee Colony

MVO: Multi-Verse Optimizer

ACO: Ant Colony Optimization

GPU: Graphics processing unit

TPU: Tensor Processing Units

GC: Google Colab

AMD: Advanced Motion Detection

HDAR: Human Detection and Activity Recognition

LRCN: Long-Term Recurrent Convolutional Network

ABSTRACT

Surveillance systems are integral part of public safety, human activity detection serving as a pivotal component for threat identification and prevention. This project proposes a novel approach integrating hybrid deep learning models. A combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are used for real-time or footage based human activity detection in surveillance. Leveraging Google Collab for model development. This system utilizes video stream coverage to capture comprehensive visual data. Through this approach, the system aims to enhance surveillance and security measures by accurately detecting suspicious activities, thus contributing to the advancement of proactive threat mitigation strategies. The methodology encompasses model design, deployment, and performance analysis, demonstrating promising results in anomaly detection and severity assessment. This research underscores the significance of advanced machine learning techniques in bolstering the efficacy of surveillance systems, with practical implications across various domains including security, smart homes, and industrial monitoring.

Chapter 1

INTRODUCTION

This chapter provides an introduction to the significance of surveillance systems in public safety, focusing on the role of Human Activity Detection for Surveillance (HAR) and the transformative impact of deep learning on automating surveillance analysis. It defines the project's aim to develop an anomaly detection system for videos using CNN and LSTM networks to identify unusual activities in surveillance footage. The chapter outlines the project's objectives, including the creation of a robust anomaly detection framework, integration of CNN and LSTM for feature extraction and temporal dependency capture, and performance evaluation using accuracy metrics. The approach involves video capture, pre-processing, and class prediction phases, with a focus on classifying student activities in a campus setting. Finally, the chapter summarizes the organization of the paper, detailing sections on related work, methodology, machine learning models, dataset description, proposed architecture, results and discussion, and references.

1.1 OVERVIEW

Surveillance systems are integral to maintaining public safety, with human activity detection serving as a fundamental component for identifying and mitigating potential threats in real-time. Leveraging the advancements in deep learning, the analysis of surveillance footage has become increasingly automated and efficient. Traditional methods of monitoring public spaces have evolved significantly, particularly in the realm of Human Activity Detection for Surveillance (HAR).

This technology has garnered significant traction in recent years, spurred by the widespread adoption of wearable devices and a growing interest in context-aware applications. Its versatility extends beyond security applications, finding utility in health monitoring, sports performance analysis, and the creation of adaptive environments responsive to human behaviour. As such, the quest for enhancing surveillance efficacy through intelligent activity detection stands at the forefront of technological innovation, promising a safer and more responsive environment for all.

Surveillance systems stand as guardians of public safety, their efficacy reliant on the ability to swiftly detect and respond to potential threats. Central to this endeavor is the discipline of Human Activity Detection for Surveillance (HAR), a dynamic field propelled by advanced algorithms and machine learning. Traditionally, monitoring public spaces demanded extensive human oversight, but with the advent of deep learning, we've witnessed a transformative shift towards automated analysis of surveillance footage. This shift is pivotal, enabling accurate and efficient monitoring of human activities in real-time, thereby bolstering threat identification and prevention measures.

Moreover, HAR's significance burgeons alongside the surge in wearable technology and the burgeoning interest in context-aware applications. Its versatility extends beyond security realms, finding practical utility in health monitoring, sports analytics, and the

creation of adaptive environments attuned to human behavior. As such, the pursuit of HAR not only fortifies surveillance systems but also fosters innovation across diverse domains, promising a future where safety and efficiency converge seamlessly.

Deep Neural Networks is one of the best architectures used to perform difficult learning tasks. Deep Learning models automatically extract features and builds high level representation of image data. This is more generic because the process of feature extraction is fully automated. From the image pixels, convolutional neural network (CNN) can learn visual patterns directly. In the case of video stream, long short-term memory (LSTM) models are capable of learning long term dependencies. LSTM network has the ability to remember things.

Surveillance systems stand as guardians of public safety, their efficacy reliant on the ability to swiftly detect and respond to potential threats. Central to this endeavour is the discipline of Human Activity Detection for Surveillance (HAR), a dynamic field propelled by advanced algorithms and machine learning. Traditionally, monitoring public spaces demanded extensive human oversight, but with the advent of deep learning, we've witnessed a transformative shift towards automated analysis of surveillance footage. The structure of the remainder of the paper is organized as follows: Section 2 delves into the related work of Human activity detection. Section 3 explicates the methodology utilized in this study. Section 4 describes the machine learning models employed. Section 5 provides an explanation of the dataset used. Sections 6 and 7 respectively discuss the proposed architecture and present the results and discussions. Lastly, Section 8 comprises the references

1.2 PROBLEM STATEMENT

Develop an anomaly detection system for videos using machine learning algorithms, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to effectively identify unusual activities or anomalies in surveillance footage.

1.3 PROJECT OBJECTIVES

1. To develop and implement a robust video anomaly detection framework capable of identifying anomalies in input video
2. Develop a combined Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network to efficiently extract spatial features and capture temporal dependencies for anomaly detection in video data.
3. Evaluate the system's performance using metrics such as Epoch and accuracy to ensure reliable anomaly detection.

1.4 APPROACH

The proposed approach will use footage obtained from camera for monitoring students' activities in a campus and send message to the corresponding authority when any suspicious event occurs.

The architecture has different phases like Video capture, Video pre-processing, Class Prediction. The system classifies the videos into three classes:

- Students fighting in campus- Suspicious class
- Walking, running- Normal class

A. Video capture

Installation of CCTV camera and monitoring the footage is the initial step in video surveillance system. Various kinds of videos are captured from different cameras, covering the whole area of surveillance.

B. Video Pre-processing

As part of pre-processing, 30 frames are extracted from each of the captured videos, frames are separated on equal time intervals. 30 extracted frames are resized to 64 x 64 and read in a numpy array of dimension (64 x 64 x 3) ~ (Image Width x Image Height x RGB) using OpenCV Library in Python.

Each Value in the frame is then Normalized by dividing it with 255.

All the 30 Normalized frames from each video are stored as sequence in numpy array with dimension 30 x 64 x 64 x 3.

C. Class Prediction

The NumPy array is given as input to the Model and the Model predicts the class of the given Video.

1.5 OUTLINE OF THE THESIS:

Chapter 2 reviews relevant literature, highlighting previous research and methodologies in human activity detection using machine learning. Chapter 3 outlines the proposed methodologies, including the block diagram, model requirements, and the creation process of the CNN-LSTM model. Chapter 4 describes the system design, focusing on the architecture and functionality of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks within the model. Chapter 5 details the implementation process, including an overview, dataset description, and testing, followed by Chapters 6, 7, 8, and 9, which present the results and conclusions, potential future enhancements, references, and any publications arising from the research

Chapter 2

LITERATURE SURVEY

This chapter provides a literature survey that explores various approaches and advancements in Human Activity Recognition (HAR) using machine and deep learning techniques. It reviews a range of studies highlighting methodologies such as sensor data analysis, metaheuristic optimization, CNN and LSTM models, and novel frameworks for HAR, emphasizing their applications, challenges, and performance metrics. This chapter sets the context for the current research by summarizing key findings and methodologies from previous work in the field of HAR, particularly focusing on surveillance and anomaly detection in video data.

1.Human activity recognition for elderly people using machine and deep learning approaches

Human Activity Recognition (HAR) has gained substantial traction in recent years, fueled by advancements in sensor technology, machine learning, and artificial intelligence. This field holds promise for various applications, from healthcare to industry, by automatically identifying and analyzing human actions using data from sensors like accelerometers and gyroscopes in smartphones and wearables. Of particular interest is its potential to support vulnerable populations, including people with disabilities and the elderly, by providing monitoring and assistance. Challenges persist, especially in developing HAR systems tailored to the unique needs of these demographics. To address these challenges, researchers are exploring multimodal features and attention-based mechanisms to enhance system performance. Machine learning and deep learning techniques, such as k-NN, Random Forest, SVM, ANN, and LSTM, play pivotal roles in HAR development, with evaluation metrics like accuracy, precision, recall, and computation speed guiding optimization efforts. The methodology involves collecting raw data, preprocessing steps like Principal Component Analysis (PCA), feature extraction, model training, and ultimately, activity detection. Despite ongoing challenges, HAR stands as a promising frontier with potential to significantly impact diverse domains and improve quality of life for individuals through robust, intelligent monitoring systems.

2. The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis

The text discusses the significance of Human Activity Recognition (HAR) applications using sensor data and the challenges associated with classifying human activities based on complex sensor datasets. It emphasizes the importance of feature selection in HAR to reduce computation time and optimize performance. The study employs various metaheuristic optimization algorithms, such as the Ant Colony Optimization (AO), Multi-Verse Optimizer (MVO), Artificial Bee Colony (ABC), and Particle Swarm Optimization (PSO), to address feature selection challenges in HAR applications. These algorithms have been widely used in diverse domains, including medical imaging, disease diagnosis, and classification tasks. By leveraging these optimization techniques, the study aims to enhance

the accuracy and efficiency of HAR systems by selecting the most relevant features from sensor data.

3. A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences.

The text discusses the evolution and challenges of human detection and activity recognition (HDAR) technologies, particularly in the context of aerial surveillance using unmanned aerial vehicles (UAVs). These technologies have applications ranging from search and rescue to law enforcement and traffic management. The paper highlights the importance of addressing challenges such as varying human scales, illumination changes, and motion blur in aerial surveillance. Various machine learning and deep learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are explored for HDAR, with a focus on extracting meaningful information from sensor and video data. The text also introduces the UCF-ARG aerial dataset, which presents unique challenges for HDAR due to its aerial perspective and small human scale. Different methods for human detection and activity recognition using this dataset are reviewed, with an emphasis on improving detection accuracy through advanced object detection algorithms like EfficientDet-D7. The paper contributes by proposing a novel HDAR system and comparing the performance of various human detection algorithms under different environmental conditions and distortions.

4.Deep CNN-LSTM with self attention model for human activity recognition using wearable sensor.

The article delves into Human Activity Recognition (HAR), addressing its significance in aiding daily life and its methods, notably video image recognition and wearable sensors. While video systems pose challenges such as cost and environmental constraints, wearable sensors offer greater flexibility and user privacy. However, despite their widespread use, the efficacy of these methods remains debatable, necessitating improvements, especially in leveraging deep learning techniques for sensor data analysis. The paper discusses the limitations of existing HAR systems, including issues related to sensor placement and device variability, and proposes a novel approach integrating Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and self-attention models for activity recognition using smartphone sensor data. The study presents the H-Activity dataset, collected from ten participants, and evaluates the proposed method against existing datasets, achieving high classification accuracy. The contributions of the paper include the development of a deep learning-based HAR framework, the creation and evaluation of the H-Activity dataset, and the reduction of dependency on traditional feature extraction methods. The article is structured to include sections on literature review, system design, experimentation, and conclusion, highlighting the importance and effectiveness of the proposed approach in advancing HAR technology.

5. Human activity recognition based on dynamic active learning.

The paper addresses challenges in human activity recognition (HAR) due to the scarcity of annotated data and the diversity of human activities. It proposes a dynamic active learning framework to minimize manual annotation efforts by selecting informative samples

iteratively. Unlike traditional active learning, this approach dynamically discovers new activities and patterns. The framework incorporates unsupervised novelty detection and clustering to identify unknown activities and employs a Best-versus-Second-Best (BvSB) sampling policy to select samples with the highest training utility. The paper introduces OCluDAL, which achieves state-of-the-art performance with reduced annotation costs and adaptability to scenarios with novel activities and patterns. Experimental evaluations on synthetic and real datasets demonstrate the effectiveness and flexibility of the proposed method. Overall, the paper contributes a comprehensive approach to HAR that considers annotation efficiency and adaptability to diverse activity patterns.

6. Abnormal Event Detection based on Analysis of Movement Information of Video Sequence

The introduction outlines the significance of abnormal event detection within video surveillance, a critical domain in computer vision research. It references various studies and methodologies employed in this area, ranging from discrete sequence analysis to probabilistic modeling using techniques such as Dynamic Bayesian Networks, Gaussian mixture models, and Hidden Markov Models (HMM). Notably, HMM emerges as a prominent tool due to its effectiveness in capturing temporal dynamics and identifying abnormal behaviors within video data. The introduction highlights the diversity of approaches and the ongoing quest to improve anomaly detection accuracy amidst challenges like noisy data and feature representation.

7. Deep Learning Approach for Suspicious Activity Detection from Surveillance Video

It underscores the ubiquitous use of video surveillance for security purposes, integrated into daily life and government initiatives like Digital India. Advantages of video surveillance, including efficient monitoring and reduced manpower, are highlighted, while challenges of manual tracking are acknowledged. The paper addresses the need for automated surveillance to enhance efficiency and accuracy in detecting abnormal events. It discusses the emergence of automated algorithms for human behavior detection, applicable in various public spaces. The integration of Artificial Intelligence, Machine Learning, and Deep Learning techniques into surveillance systems is emphasized for feature extraction and classification. Computer vision methods are vital for public safety, involving environment modeling, motion detection, and behavior understanding. Deep Neural Networks, particularly Convolutional Neural Networks and Long Short-Term Memory models, are recognized for their effectiveness in visual pattern recognition and long-term dependency learning. The proposed system aims to use CCTV footage to monitor human behavior on campuses, issuing alerts for suspicious activities. Overall, the introduction highlights the importance of automated video surveillance systems in ensuring public safety and security, addressing challenges in human behavior recognition.

8. Inspection of suspicious human activity in the crowdsourced areas captured in surveillance areas captured in surveillance cameras

The introduction outlines the focus of the research within the realm of security techniques, particularly in addressing challenges related to repeated aim detection applications. It highlights the aim of computerized surveillance systems in supporting human operators by

autonomously detecting objects and analyzing their actions through computer vision, pattern recognition, and signal processing techniques. Despite advancements in these fields, the paper acknowledges the persistent difficulty in achieving a practical end-to-end surveillance system due to real-world challenges. However, with the evolution of computing technology, the adoption of multi-camera and multi-modal structures has become feasible, catering to the needs of efficient surveillance systems in various security applications. Visual surveillance is identified as a crucial area in pattern analysis and machine intelligence, crucial for aiding intelligence and law enforcement agencies in combating crime. The primary objective of visual surveillance systems is to detect irregular object behaviors and raise alarms accordingly, often utilizing Advanced Motion Detection (AMD) algorithms. Following the detection of moving objects, object categorization becomes imperative for interpreting their movements and behaviors in context, thus serving as a vital component of a comprehensive visual surveillance system.

9. LSTM-CNN architecture for human activity recognition.

Human activity recognition (HAR) is pivotal in daily life, enabling the extraction of profound insights from sensor data. It has found applications in home behavior analysis, video surveillance, gait analysis, and gesture recognition. With advancements in sensor and computing technologies, sensor-based HAR has gained traction, categorized into fixed and mobile sensor approaches. Fixed sensors include cameras and acoustic sensors, while mobile sensors involve accelerometers and gyroscopes. While fixed sensor methods offer better accuracy, they are limited by privacy concerns and environmental factors. Mobile sensor-based approaches offer portability and have seen extensive research for ubiquitous activity recognition. The paper reviews current sensor-based HAR research, discusses dataset descriptions and preprocessing, presents a proposed LSTM-CNN architecture, and evaluates experimental results, considering network structure and hyperparameters. Finally, it provides a summary of the research findings.

10. Human Activity Recognition via Hybrid Deep Learning Based Model

The text discusses the development of a hybrid model for human activity recognition (HAR) using a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The goal is to accurately recognize different types of physical activities performed by individuals, especially in an indoor environment. The text also highlights the generation of a new dataset comprising 12 different physical activities performed by 20 participants using the Kinect V2 sensor. The results indicate that the hybrid CNN-LSTM model achieved a high accuracy of 90.89% on a 30 frames sequence, outperforming other deep learning approaches in activity recognition. Additionally, the text mentions the potential limitations of the proposed model in recognizing the activity of multiple people simultaneously and discusses future research directions to expand the dataset and improve the model's performance.

11. Direction-Independent Human Activity Recognition Using a Distributed MIMO Radar System and Deep Learning.

The text discusses the development of a direction-independent Human Activity Recognition (HAR) system using a multiperspective 2x2 distributed MIMO radar

configuration. It goes on to describe the design and development of the conventional SISO radar-based HAR system, SISO (1-D), and the direction-independent SISO radar-based HAR system, SISO (2-D). The classification performance of these systems is evaluated using the HAR dataset. The text also discusses the feature extraction network and deep convolutional neural network (DCNN) used to classify human activities. Finally, it presents a detailed analysis of the classification performance of the developed HAR systems

12. Recognition of Daily Human Activity Using an Artificial Neural Network and Smartwatch

The research presents a human activity recognition (HAR) system that utilizes a smartwatch and artificial neural network (ANN) to accurately classify 11 different activities with a 95% accuracy rate. The smartwatch collects acceleration data, which is then transmitted to a smartphone and subsequently to a server for processing and classification. The system also incorporates location information to enhance performance. Potential real-world applications include energy-efficient activity prediction and enhanced convenience through automated actions based on recognized activities.

13. Human Activity Recognition with Smartphone and Wearable Sensors Using Deep Learning Techniques

The paper introduces Human Activity Recognition (HAR) as a field that utilizes sensors in smartphones and wearable devices to infer human activities from raw time-series signals. HAR is particularly significant in smart home environments for monitoring human behaviors in ambient assisted living. The system comprises modules for data acquisition, pre-processing, feature extraction, selection, and classification. While traditional Machine Learning classifiers have been employed with rustic feature extraction processes, recent advancements leverage Deep Learning techniques for more efficient feature retrieval and classification. The review paper aims to provide an overview of these Deep Learning techniques applied in smartphone and wearable sensor-based recognition systems. It categorizes the techniques into conventional and hybrid models, discussing their merits, limitations, and benchmark datasets used. Additionally, the paper highlights challenges and areas for future research and improvement within the field of HAR.

14. Enhanced Human Action Recognition Using Fusion of Skeletal Joint Dynamics and Structural Features

The introduction highlights the exponential growth in video data generated by devices like smartphones and CCTV cameras, emphasizing the importance of extracting valuable insights from this data. Human action recognition (HAR) is identified as a crucial research area with applications ranging from surveillance systems to sports video analysis. HAR involves detecting and understanding human behaviors from video sequences, requiring a combination of computer vision and pattern recognition techniques. The terminology used in HAR literature, such as activity, behavior, and gesture, is discussed, noting their interchangeability. Challenges in HAR include intraclass variation and interclass similarity, which make distinguishing between activities difficult. Two main approaches for HAR,

based on global and local descriptors, are outlined, with recent emphasis on skeleton-based approaches due to the availability of depth sensors. Various datasets for evaluating action recognition algorithms are mentioned, differing in factors like the number of classes and complexity of actions. The paper's contributions include proposing a method for action recognition based on joint angle information and neural network-based score-level fusion for classification. The structure of the paper, including sections on existing techniques, the proposed approach, experimental results, and conclusions, is outlined.

15. Outlier Detection in Wearable Sensor Data for Human Activity Recognition (HAR) Based on DRNNs

This paper introduces a novel algorithm that combines outlier detection and human activity recognition (HAR) using wearable sensor data. It aims to detect secondary activities within main activities and extract data segments of specific sub-activities from different activities. Previous studies have utilized machine learning algorithms for HAR, with deep recurrent neural networks (DRNNs) proving optimal for wearable sensor data analysis. The proposed DRNN-based algorithm is designed for outlier detection in HAR and validated for both intra- and inter-subject cases, as well as for sub-activity recognition, using two datasets. The first dataset, involving 15 users and four major activities, demonstrates successful detection of outliers in walking activities and extraction of walking segments from other activities. The second dataset, featuring four users and different settings and sensors, evaluates the generalization of results. Overall, the algorithm shows promising results in outlier detection and sub-activity recognition across different scenarios and datasets

16. Orientation-Independent Human Activity Recognition Using Complementary Radio Frequency Sensing

This paper introduces a novel RF sensing approach for detecting accidental falls and recognizing human activities, addressing limitations of existing single monostatic radar-based systems. Utilizing a distributed mmWave MIMO radar system with two separate monostatic radars positioned orthogonally indoors, it captures motion from different aspect angles to generate micro-Doppler signatures. Mean Doppler shifts (MDSs) are computed from these signatures, and statistical, time-, and frequency-domain features are extracted. Feature-level fusion techniques are employed to combine these features, with classification carried out using a support vector machine. Evaluation was conducted on an orientation-independent human activity dataset from six volunteers, comprising over 1350 trials of five activities performed in different orientations. The proposed approach achieved an overall classification accuracy ranging from 98.31% to 98.54%, overcoming limitations of single monostatic radar-based systems and outperforming them by 6%.

Chapter-3

PROPOSED METHODOLOGY

This chapter details the proposed methodology for human activity detection in surveillance using CNN and LSTM models. The approach leverages CNNs for extracting spatial features from video frames and LSTMs for capturing temporal dependencies, enabling accurate activity recognition. This methodology is evaluated using the UCF50 dataset, with performance measured through accuracy, precision, and recall metrics to ensure robust detection and classification of activities in video surveillance scenarios.

In the methodology of human activity detection for surveillance using CNN + LSTM models, the approach integrates the strengths of both convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to effectively analyse video data from surveillance cameras. Initially, raw video data captured from two cameras is fed into the CNN model. The CNN processes this visual data, extracting meaningful visual features that represent different aspects of human activities such as motion patterns, spatial relationships, and object interactions. These extracted visual features serve as a rich representation of the input video frames. Subsequently, the output of the CNN, consisting of visual features, is passed on to the LSTM network.

The LSTM, known for its capability in capturing temporal dependencies and sequential patterns, performs sequence learning on the visual features obtained from the CNN. By analysing the temporal evolution of visual features over time, the LSTM model can discern complex activity patterns and dynamics inherent in the surveillance footage. As the LSTM processes the sequential data, it learns to recognize patterns indicative of various human activities, such as walking, running, gesturing, or interacting with objects. The LSTM's ability to retain information over long sequences enables it to capture the temporal context crucial for accurate activity detection. Finally, the output of the LSTM model represents the detected human activities based on the learned patterns and sequences from the input video data. These detected activities can include a wide range of behaviours and actions observed in the surveillance footage, providing valuable insights for security monitoring, anomaly detection, or behavioural analysis purposes.

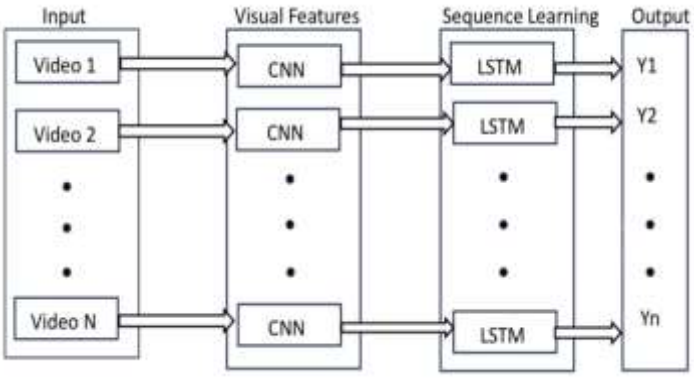

Fig.3.1 CNN and LSTM Model

We employ widely known performance evaluation criteria, namely, accuracy, precision, and recall, to measure the recognition performance of the proposed system. Accuracy measures the total percentage of the accurate recognition rate of the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100$$

Here, TP classifies the positive class as positive, FP classifies the positive class as negative, TN classifies the negative class as negative, and FN classifies the negative class as positive.

Overall, the combined CNN + LSTM approach leverages the complementary strengths of both architectures, enabling robust and accurate human activity detection in surveillance scenarios by effectively capturing both spatial and temporal dynamics inherent in video data.

3.1 BLOCK DIAGRAM

In the block diagram for human activity detection in surveillance, the process begins with the input of video data captured by cameras. This raw video data serves as the primary source of information for the subsequent analysis. The video data is then fed into a convolutional neural network (CNN), which specializes in extracting visual features from images or frames. The CNN processes each frame of the video, extracting relevant visual patterns, motion cues, and spatial relationships that characterize different human activities. he output of the CNN, which consists of the extracted visual features, is then passed on to a long short-term memory (LSTM) network.

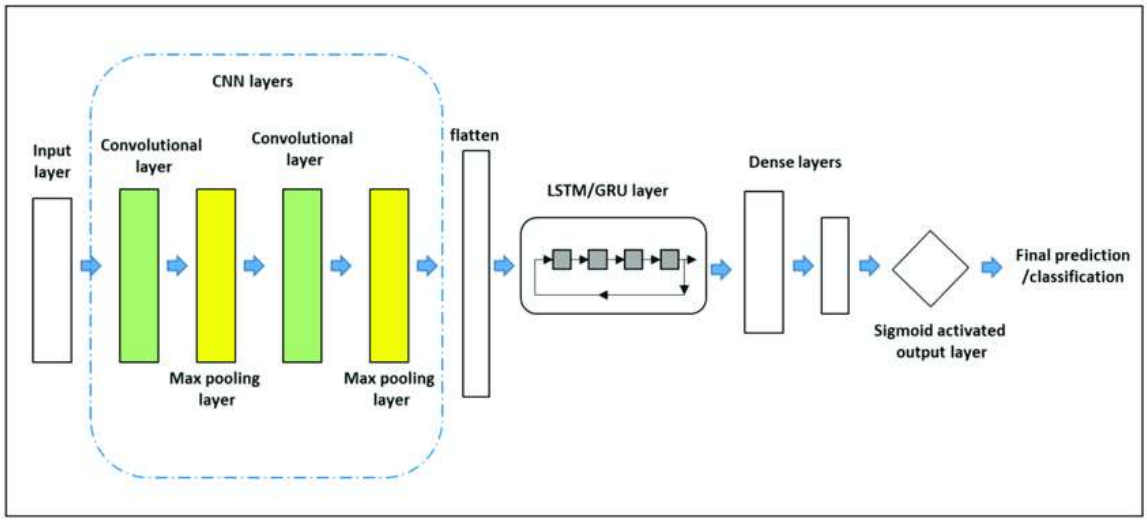

Fig.3.2 Block Diagram of the proposed model for activity detection

Unlike traditional neural networks, LSTM networks are capable of capturing temporal dependencies and sequential patterns in data. In this context, the LSTM analyzes the sequence of visual features over time, learning the dynamic evolution of human activities within the video footage. Once the CNN and LSTM models have processed the video data and extracted relevant features, the combined model is trained using labeled data. During the training phase, the model learns to associate specific visual patterns and temporal sequences with corresponding human activities. Through iterative adjustments of model parameters, such as weights and biases, the model improves its ability to accurately classify different activities based on the learned features. After the training phase, the model is ready for inference. New, unseen video data can be input into the trained model, which then applies the learned classification rules to predict the human activities present in the footage. The model's output provides classifications or labels for each segment of the video, indicating the detected activities such as walking, running, standing, or interacting with objects.

In summary, the block diagram illustrates a pipeline for human activity detection in surveillance, leveraging a combination of CNN and LSTM models. By processing video data through these specialized architectures, the model can effectively capture both spatial and temporal characteristics of human activities, enabling accurate and robust detection and classification in real-world surveillance scenarios.

3.2 MODEL REQUIREMENTS

3.2.1 Google Colab

Google Colab is a cloud-based platform that allows users to run Python code and execute machine learning tasks in a collaborative environment. It provides access to powerful GPU and TPU resources for accelerated computation without the need for expensive hardware. Users can write and execute code in a Jupyter Notebook-style interface, which supports various libraries and frameworks commonly used in machine learning, including TensorFlow and PyTorch. Colab notebooks are stored on Google Drive, making it easy to share and collaborate on projects with team members. Additionally, Colab offers pre-installed libraries and dependencies, simplifying setup and configuration for machine learning projects. By leveraging Google Colab, users can efficiently train and evaluate CNN-LSTM models for anomaly detection in video data, benefiting from its convenience, scalability, and cost-effectiveness.

3.2.2 DATASET : UCF50

The UCF50 dataset is a comprehensive collection of videos designed for human activity recognition and detection tasks. It comprises 50 distinct action categories, each containing at least 100 video clips, resulting in a total of 6,618 video clips. These categories encompass a wide range of activities, including sports, household tasks, and leisure pursuits, among others. The dataset offers diversity in scenes, actors, and camera motions as the videos are sourced from YouTube. This diversity ensures that models trained on the dataset are robust and generalize well across various scenarios. Each video clip is accompanied by annotations specifying the action category, providing valuable ground truth for model training and evaluation. Videos are available in different resolutions and durations, adding to the dataset's versatility. The UCF50 dataset is frequently utilized as a benchmark for evaluating the performance of machine learning algorithms in human activity recognition and video detection tasks. Researchers and practitioners can access the dataset for research purposes, promoting collaboration and innovation in the field of computer vision. Its accessibility and richness in content make it an invaluable resource for advancing the capabilities of video detection systems.

3.2.3 Model Descriptions:

3.2.3.1 CNN Model:

Spatial Feature Extraction:

- The CNN component processes individual frames of the video sequence.
- Convolutional layers detect spatial patterns such as edges, textures, and object features.
- Pooling layers reduce spatial dimensions and retain important features.

Feature Maps:

- CNN generates feature maps representing spatial information extracted from each frame.
- These feature maps capture relevant visual patterns and details within the video frames.

3.2.3.2 LSTM Model:

Temporal Modeling:

- The LSTM network processes the sequence of feature maps generated by the CNN over time.
- It captures temporal dependencies and sequential patterns in the video data.

Memory Cells:

- LSTM units contain memory cells that maintain information over long sequences.
- Gates within LSTM units regulate the flow of information, allowing the network to learn and forget information selectively.

3.2.3.3 Train & Test Split Data

To effectively train and test a CNN and LSTM model for video detection, 75% of the dataset is allocated for training and 25% for testing, ensuring a sufficient amount of data for both phases while maintaining a reasonable split. During training, the model learns to extract spatial features using the CNN component and capture temporal dependencies with the LSTM component, optimizing its parameters to minimize the training loss. The remaining 25% of the data, unseen during training, is utilized for testing, allowing for an unbiased evaluation of the model's performance on unseen examples. This separation helps assess the model's ability to generalize to new data and detect anomalies accurately in real-world scenarios. By training on a majority of the data and testing on a distinct subset, we can gauge the model's efficacy in detecting anomalies while avoiding overfitting to the training data. The training-testing split ensures robustness and reliability in evaluating the model's performance metrics, such as accuracy, precision, recall, and F1-score, providing insights into its effectiveness for video detection tasks.

3.3 MODEL CREATION

A deep learning network, LRCN is using in our proposed system for suspicious activity detection from video surveillance.

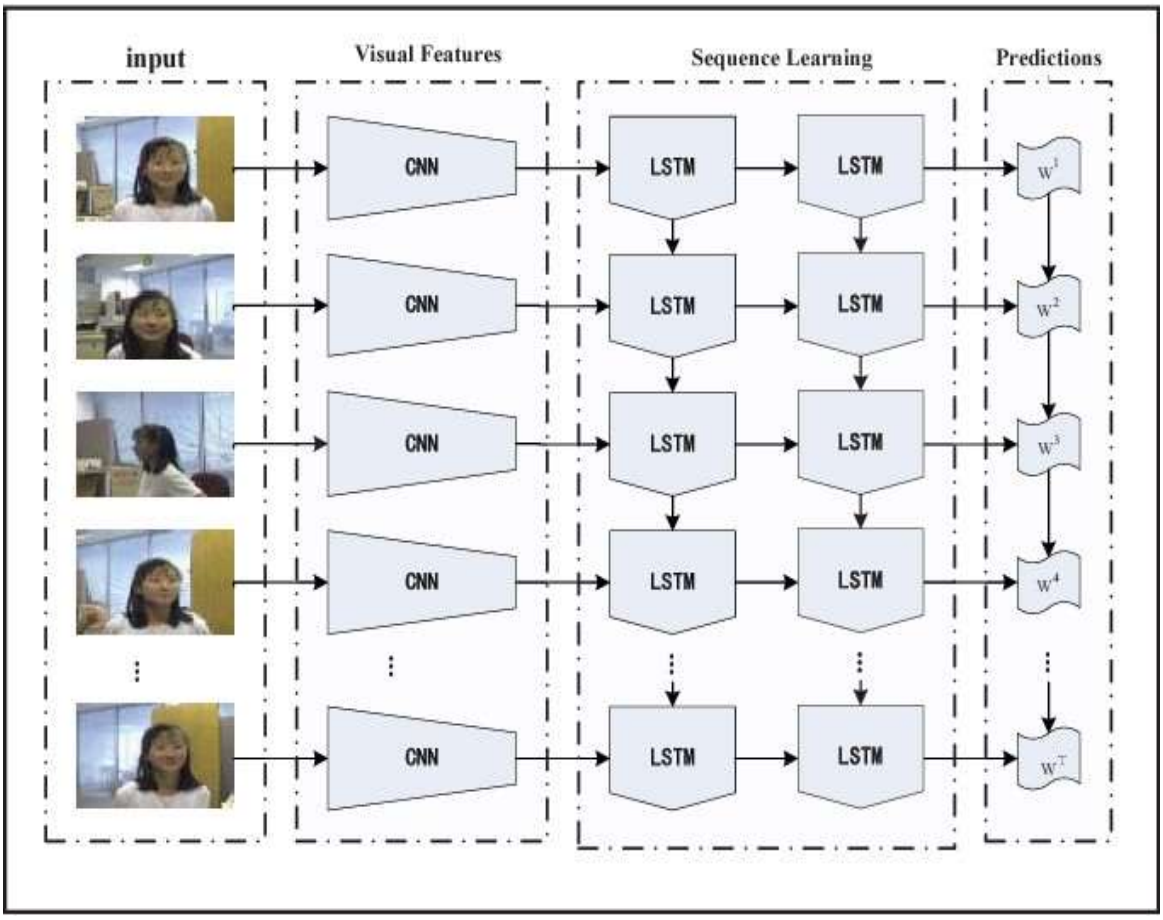

Fig.3.3 LRCN Model

In 2016 a group of authors suggested end-to-end trainable class of architectures for visual recognition and description. The main idea behind LRCN is to use a combination of CNNs to learn visual features from video frames and LSTMs to transform a sequence of image embeddings into a class label, sentence, probabilities, or whatever you need. Thus, raw visual input is processed with a CNN, whose outputs are fed into a stack of recurrent sequence models.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

Chapter-4

SYSTEM DESIGN

This Chapter details the system design for human activity detection in surveillance using a CNN and LSTM hybrid model. The CNN architecture focuses on feature extraction from video frames, while the LSTM captures temporal dependencies, forming an effective pipeline for activity recognition. The model is trained on the UCF50 dataset, with performance evaluated through metrics such as accuracy, precision, and recall to ensure robust detection and classification of human activities in real-world surveillance scenarios.

4.1 Convolution Neural Network (CNN)

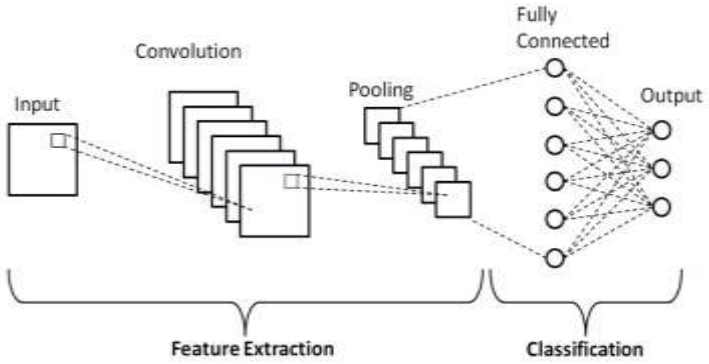

Fig. 4.1: Convolution Neural Network Architecture

In human activity detection for surveillance, CNN (Convolutional Neural Network) models are utilized for their efficacy in analysing visual data. The block diagram of a CNN model for this purpose typically consists of three main steps: input, feature extraction, and classification. At the input stage, raw image data or video frames are fed into the CNN model. These images serve as the primary source of information for the network to analyse.

The feature extraction phase, often referred to as the convolution and pooling layers, involves the application of filters to the input images. Convolutional layers detect various features such as edges, textures, and patterns within the input data. Pooling layers then down sample the feature maps produced by convolution, reducing their dimensionality while retaining essential information. Together, these layers effectively extract meaningful features from the input images, capturing relevant visual patterns that are crucial for human activity detection. Following feature extraction, the fully connected layers come into play for classification. These layers take the high-level features extracted by the convolution and pooling layers and map them to specific classes or categories of human activities. Through the process of training, the CNN learns to distinguish between different activities based on the extracted features, enabling accurate classification of observed behaviours in surveillance footage.

In summary, the block diagram of a CNN model for human activity detection comprises input, feature extraction (consisting of convolution and pooling layers), and classification (implemented through fully connected layers). This architecture effectively leverages the power of machine learning to analyse visual data and identify human activities in surveillance scenarios.

4.2 Long Short-Term Memory (LSTM)

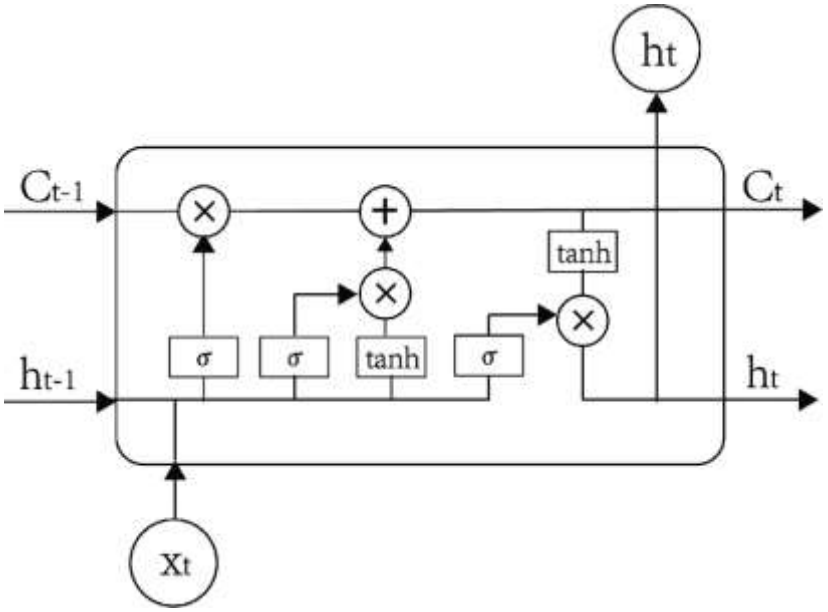

Fig.4.2 Single long short-term memory (LSTM) cell.

In the LSTM architecture, each LSTM cell incorporates three crucial gates: forget, input, and output gates. These gates regulate the flow of information within the cell state, which serves as a memory unit, enabling the LSTM to store and propagate information across time steps. Figure 1 illustrates the interconnections between these gates and the cell state. The forget gate, which receives input from both the current time step $XtXt$ and the previous output $ht-1ht-1$, utilizes a sigmoid activation function to determine which information to retain or discard. If the output of the sigmoid function is 1, the corresponding information is preserved, while an output of 0 signifies complete removal. Equation (1) demonstrates the computation of the forget gate. Subsequently, the input gate decides what new information from the current input (Xt, $ht-1$)(Xt,$ht-1$) should be added to the cell state.

This information is combined with a new candidate vector $C\sim tC\sim t$, generated by the tanh activation function, to produce an updated cell state $CtCt$. Equations (2)-(4) detail the calculations involved in determining the input gate, new candidate values, and cell state, respectively. Following this, the output gate is determined based on filtered information, employing two different activation functions, and specifies the next hidden state. The previous hidden state $ht-1ht-1$ and the current input $xtxt$ are first passed through a sigmoid activation function, while the updated cell state is fed into a tanh activation function. The outputs of these functions are then multiplied to generate the next hidden state. Equations (5) and (6) outline the computation of the output gate and hidden state, respectively.

In essence, the LSTM cell operates as a memory unit, selectively erasing, reading, and writing information based on the decisions made by the forget, input, and output gates. This mechanism enables LSTMs to effectively capture long-term dependencies and relationships in sequential data.

where x is the input data, σ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, W is the weight matrix.

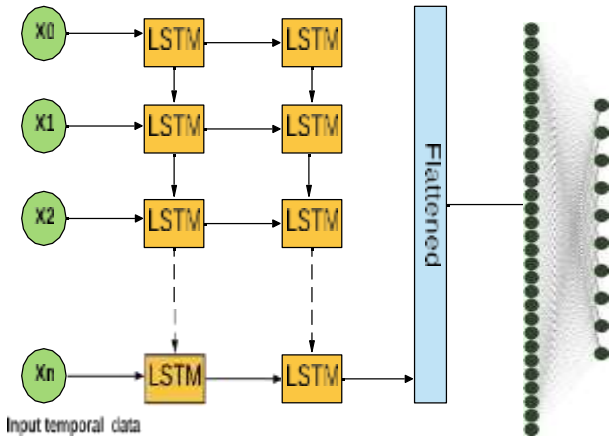

Fig.4.3 Architecture of the LSTM model for human activity recognition.

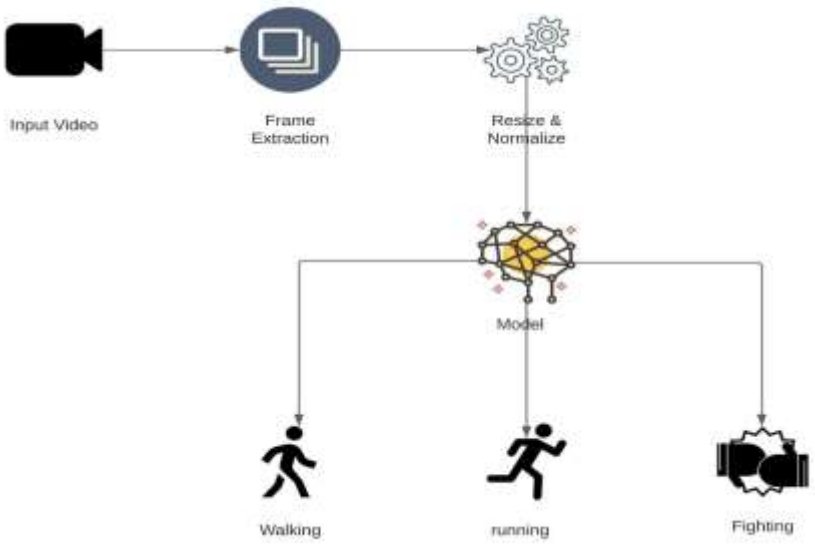

Fig.4.4 Proposed Model

4.3 MODEL

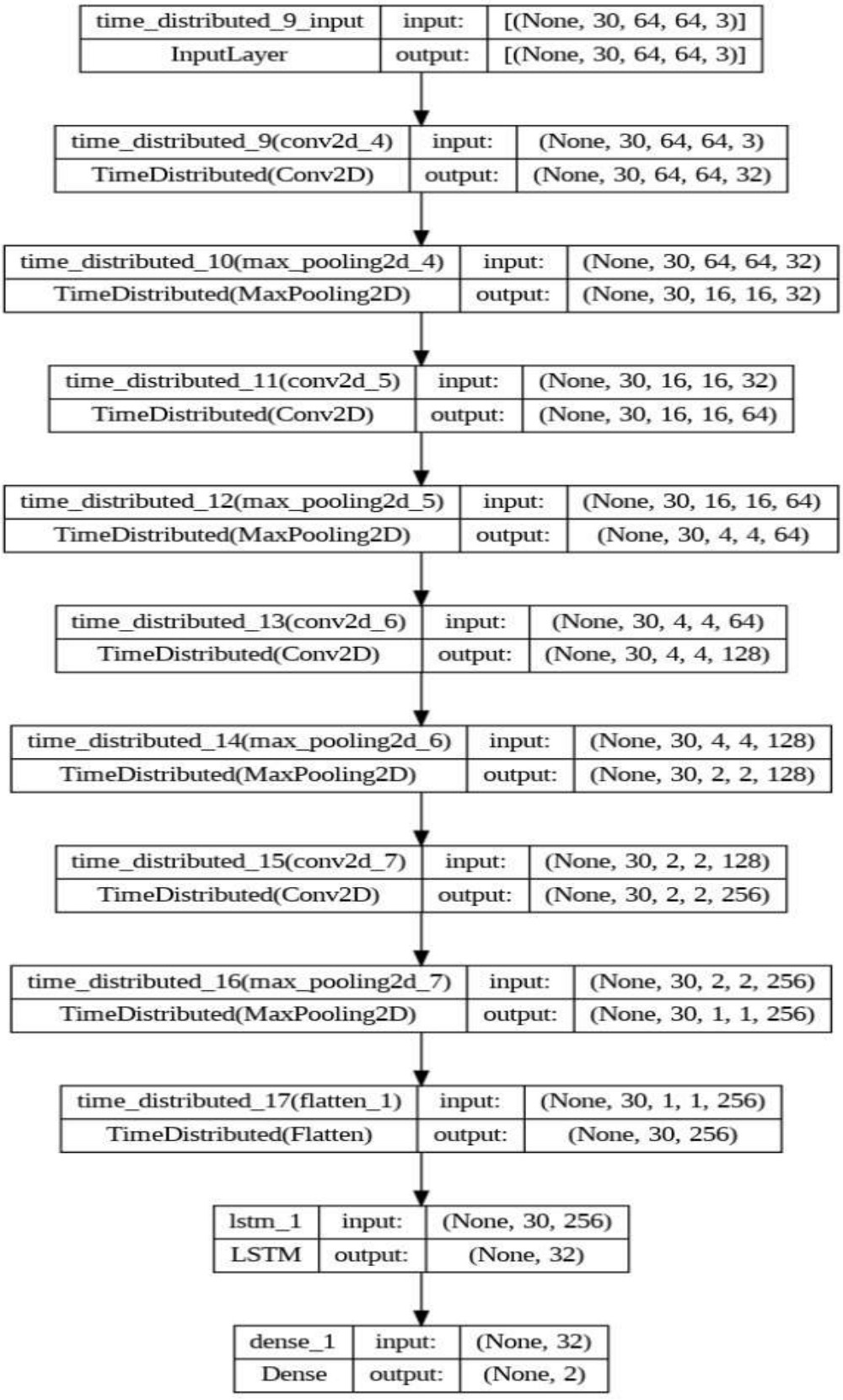

Fig.4.5 Model Plot

The neural network described is a convolutional neural network (CNN) designed for processing sequential data, particularly suited for tasks such as video analysis or sequential prediction. Let's break down its architecture and functionality:

1. **Input Layer:**
    - The network accepts a sequence of 30 frames, where each frame has a spatial resolution of 64x64 pixels and three color channels (RGB).
2. **Convolutional Layers (9 Layers):**
    - The first 9 layers are convolutional layers responsible for extracting spatial features from each frame in the input sequence.
    - Each convolutional layer applies a set of learnable filters to the input feature maps, capturing spatial patterns at different scales.
3. **Pooling Layers (4 Layers):**
    - Following the convolutional layers, there are 4 pooling layers, typically max-pooling layers.
    - Pooling layers downsample the spatial dimensions of the feature maps, reducing computational complexity while preserving important features.
4. **Flattening Layer:**
    - After the pooling layers, there is a flattening layer that transforms the 3D feature maps into a 1D vector, ready for input to the fully connected layers.
5. **Fully Connected Layers (4 Layers):**
    - The last 4 layers are fully connected layers, also known as dense layers.
    - These layers integrate the spatial features extracted by the convolutional layers and capture high-level patterns and relationships in the data.
6. **Output Layer:**
    - The output of the network is a sequence of 30 vectors, each containing 2 elements.
    - Each element in the output vector represents a prediction or classification score for a specific aspect of the input data.

Overall, this network architecture is tailored for processing sequential data, such as video frames, by hierarchically extracting spatial features using convolutional layers and capturing temporal dependencies through subsequent layers. The output sequence of vectors enables the network to make predictions or classifications for each frame in the input sequence, making it suitable for tasks like video action recognition or anomaly detection.

4.3.1 Model Training

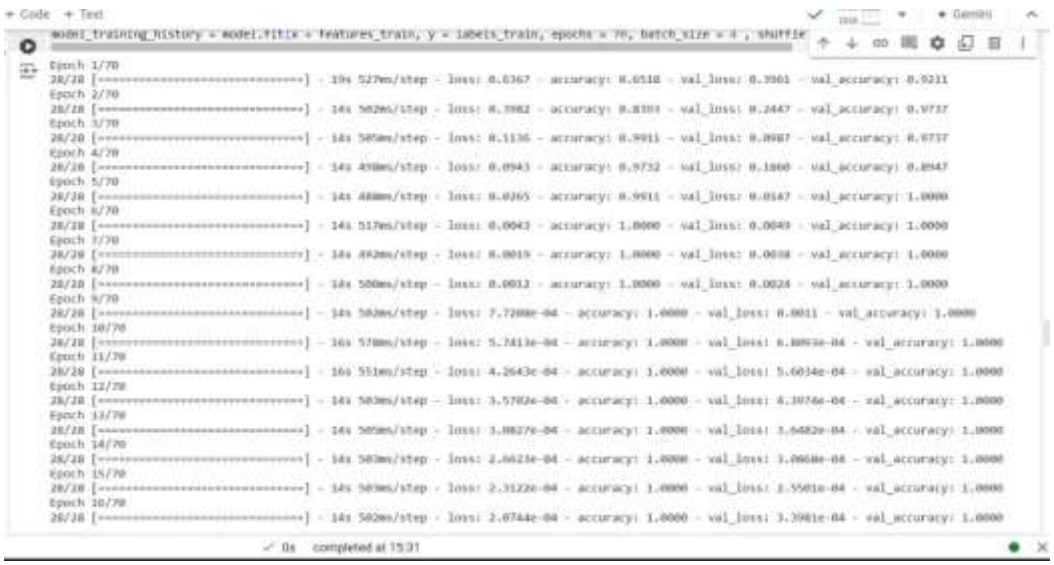

Fig.4.6 Accuracy & Epoch

In this training scenario, the model is being trained on a dataset consisting of 70 features and 4 labels, implying a classification task where the model learns to map input features to one of the four predefined labels. The training process spans 70 epochs, indicating that the entire dataset will be iterated over 70 times during training. Each epoch processes data in batches of 4 samples, enhancing computational efficiency and stability.

Setting the shuffle parameter to True ensures that the data is shuffled before each epoch, preventing the model from memorizing the order of the samples and improving generalization. Shuffling the data before each epoch helps the model learn to generalize better to unseen examples.

During training, the model's progress is monitored through the output, which displays the loss and accuracy metrics for each epoch. The loss metric quantifies the disparity between the model's predictions and the true labels, serving as a measure of how well the model is performing. On the other hand, the accuracy metric reflects the proportion of correctly classified samples out of the total.

Saving the model after every epoch to a file named "model.h5" ensures that the most recent version of the model is preserved, allowing for easy retrieval and continuation of training if necessary. This practice also helps prevent loss of progress in case of unexpected interruptions or failures during training.

Finally, achieving 100% accuracy at the end of training suggests that the model has successfully learned to classify all samples in the dataset correctly. However, it's essential to cautiously interpret this result, as it could indicate overfitting if the model performs poorly on unseen data. Further evaluation, such as using a separate validation dataset, is recommended to validate the model's generalization capability.

Chapter-5

IMPLEMENTATIONS

The objective of this chapter is to provide the implementation to detect human activities in surveillance videos using a combination of CNN and LSTM networks. The UCF50 dataset, containing 50 action categories, is pre-processed by splitting into training, validation, and testing sets, extracting frames, resizing to 224x224 pixels, and normalizing. The model architecture consists of a pre-trained CNN for spatial feature extraction and an LSTM for temporal sequence modelling, trained using categorical cross-entropy loss and the Adam optimizer. Model evaluation includes accuracy, precision, recall, F1-score, and a confusion matrix. Testing involves preparing a test dataset similarly to the training data, calculating key metrics, and generating a detailed classification report.

5.1 Overview

The objective of this implementation is to detect human activities in surveillance videos using a combination of CNN and LSTM networks. The UCF50 dataset, which contains 50 different action categories, is utilized for training and evaluation.

Certainly! Here is the implementation description in sentence form:

The paper titled "Human Activity Detection for Surveillance" implements a model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to detect human activities in surveillance videos using the UCF50 dataset. The UCF50 dataset, containing 50 different action categories, is first pre-processed by downloading and splitting it into training, validation, and testing sets. Frames are extracted from each video at a consistent frame rate, resized to 224x224 pixels, and normalized to the range [0, 1].

The model architecture consists of two main parts: a CNN for spatial feature extraction and an LSTM for temporal sequence modelling. A pre-trained CNN model, such as VGG16 or ResNet50, is used as the backbone for feature extraction, where the top fully connected layers are removed to obtain feature maps from the last convolutional layer. These feature maps, serving as a sequence of feature vectors, are then passed to an LSTM network. The LSTM network comprises one or more LSTM layers with a specified number of units, followed by a dropout layer to prevent overfitting, and a fully connected layer with SoftMax activation for classification.

For training, the model uses the categorical cross-entropy loss function and the Adam or RMSprop optimizer, with a starting learning rate of 0.001 and a scheduler to reduce it on plateau. An appropriate batch size, such as 16 or 32, is chosen based on GPU memory, and the model is trained for 50-100 epochs with early stopping based on validation loss.

The evaluation of the model includes metrics such as accuracy, precision, recall, and F1-score, along with a confusion matrix to visualize the performance across different activity classes. Cross-validation is employed to ensure robustness.

The implementation begins with importing necessary libraries like TensorFlow/Keras, NumPy, and OpenCV. The data preprocessing step involves a function to extract and preprocess frames from videos. The CNN-LSTM model is defined using a pre-trained CNN model and sequentially adding LSTM and Dense layers. The model is then compiled and trained using early stopping and learning rate reduction callbacks to optimize performance. Finally, the model's performance is evaluated using classification reports and confusion matrices, visualized using tools like seaborn and matplotlib.

5.2 DATASET DESCRIPTION

The UCF50 dataset is a widely used benchmark for human activity recognition, featuring 6,680 realistic action videos categorized into 50 diverse action classes. These categories encompass a range of activities such as sports (e.g., basketball, soccer juggling), musical instruments (e.g., playing guitar, playing piano), body movements (e.g., walking, running), and human-object interactions (e.g., brushing teeth, blowing candles). The videos, sourced from YouTube, vary in resolution and length, adding to the dataset's complexity and realism. Each video is labelled with a single action category and organized into groups to evaluate cross-subject and cross-scenario generalization. The UCF50 dataset is publicly available for download and serves as a standard benchmark in research areas like human activity recognition, video classification, and action detection.

Fig.5.1 UCF50 Dataset

The UCF50 dataset includes images depicting a wide variety of human actions across different settings. These actions are categorized into 50 distinct groups, encompassing activities such as:

Sports: Examples include playing basketball, tennis, soccer, and golf, with images showing various dynamic movements associated with these sports.

Exercises: Depictions of physical exercises like push-ups, pull-ups, yoga, and aerobics, showcasing different postures and forms.

Daily Activities: Everyday actions such as brushing teeth, walking the dog, riding a bike, and playing musical instruments, providing a glimpse into routine life.

Each image captures the essence of the action, offering a rich visual dataset for studying and analysing human motion and activities.

5.3 TESTING

1. Test Dataset Preparation

The test dataset, which is a subset of the UCF50 dataset not seen during training, is prepared in the same manner as the training data. This includes:

- Extracting frames from test videos at a consistent frame rate.
- Resizing frames to a fixed size, such as 224x224 pixels.
- Normalizing pixel values to the range [0, 1].

2. Model Evaluation

The trained CNN-LSTM model is evaluated on the test dataset using the following metrics:

Accuracy: The overall accuracy is calculated as the percentage of correctly predicted activities out of all test samples. This metric provides a general measure of the model's performance.

Precision: Precision is calculated for each activity class as the ratio of true positive predictions to the sum of true positive and false positive predictions. This metric indicates how many of the predicted positive instances are actually correct.

Total Loss: The total loss on the test dataset is computed using the same loss function used during training, typically categorical cross-entropy. This metric indicates how well the model's predicted probabilities align with the actual labels.

Total Accuracy: This is essentially the same as overall accuracy, providing a summary of the model's ability to correctly predict the activities in the test set.

3. Classification Report

A detailed classification report is generated to provide precision, recall, and F1-score for each activity class, offering a comprehensive view of the model's performance across all categories. During the testing phase, the model's performance is rigorously evaluated on the UCF50 test dataset. Key metrics such as accuracy, precision, total loss, and total accuracy are calculated to assess the model’s effectiveness in detecting human activities. Additionally, a confusion matrix and a detailed classification report are generated to provide further insights into the model's performance across different activity classes.

Chapter 6

RESULTS AND CONCLUSION

This Chapter presents the results and conclusion of the CNN-LSTM approach in human activity recognition. Graphs titled "Total Loss vs Total Validation Loss" and "Total Accuracy vs Total Validation Accuracy" showcase the model's performance over 15 epochs, illustrating significant improvement and stabilization at low loss values and high accuracy levels, respectively. Additionally, an image captures the CNN-LSTM model training process in Google Colab, providing insights into batch and epoch information along with progress visualization through a progress bar. These results affirm the effectiveness of the proposed approach in accurately detecting human activities in surveillance videos.

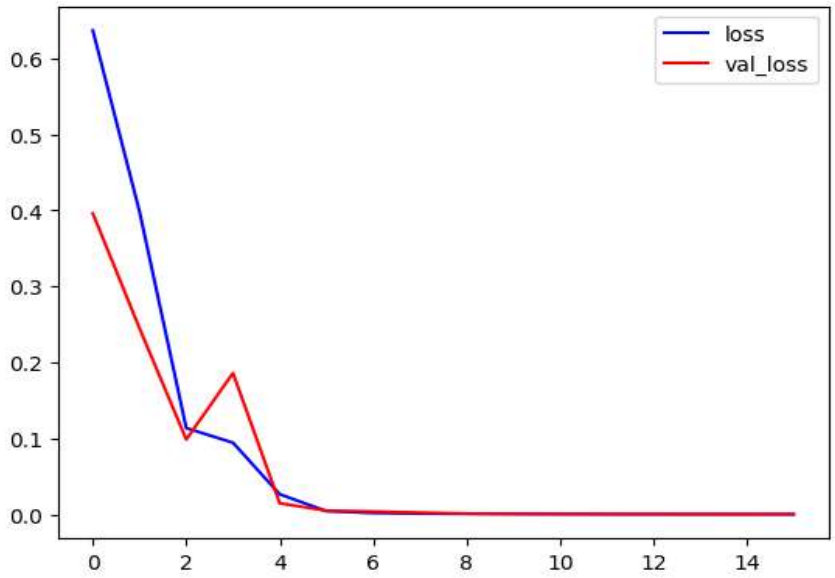

Fig.6.1 Total Loss vs Total Accuracy

The graph titled "Total Loss vs Total Validation Loss" depicts the training and validation losses of a model over 15 epochs. The x-axis represents the number of epochs, ranging from 0 to 15, while the y-axis represents the loss values, ranging from 0.0 to 0.6.

Two lines are plotted on the graph:

- The blue line represents the training loss (loss).
- The red line represents the validation loss (val_loss).

Key observations:

- **Initial Losses**: At epoch 0, the training loss is relatively high, starting above 0.6, while the validation loss is around 0.4.

- **Rapid Decrease**: Both losses decrease rapidly in the first few epochs. By epoch 2, the training loss drops below 0.1, and the validation loss also decreases significantly.
- **Minor Fluctuation**: Around epoch 3, there is a slight increase in the validation loss, indicating a minor fluctuation, but it quickly stabilizes.
- **Convergence**: From epoch 4 onwards, both the training loss and validation loss converge and remain very low, close to 0.0. This indicates that the model has effectively minimized both losses.

Overall, the graph demonstrates significant improvement in the model's performance over the epochs. Both training and validation losses decrease and stabilize at low values, suggesting effective training and good generalization

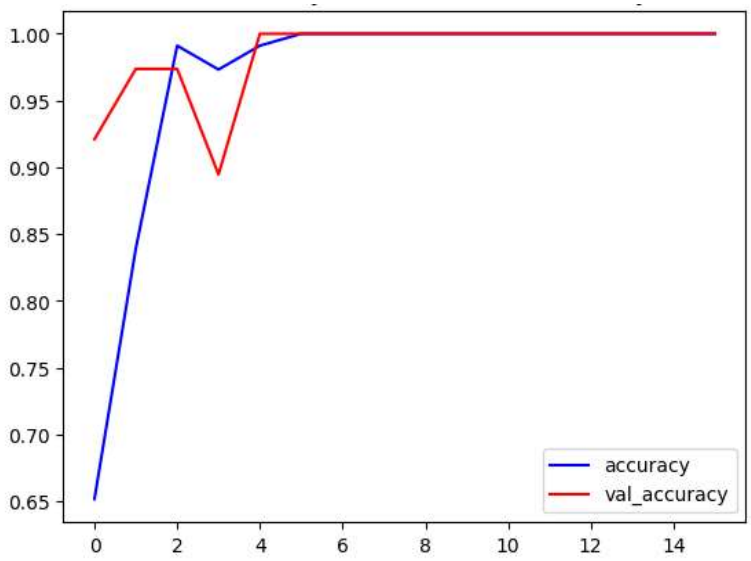

Fig.6.2 Total Accuracy vs Total Validation Accuracy

The graph titled "Total Accuracy vs Total Validation Accuracy" illustrates the training and validation accuracies of a model over 15 epochs. The x-axis represents the number of epochs, ranging from 0 to 15, while the y-axis represents the accuracy values, ranging from 0.65 to 1.00.

Two lines are plotted on the graph:

- The blue line represents the training accuracy (accuracy).
- The red line represents the validation accuracy (val_accuracy).

Key observations:

1. **Initial Accuracy**:
    - At epoch 0, the training accuracy starts around 0.65, while the validation accuracy is higher, starting around 0.95.

2. **Rapid Initial Increase**:
    - Both accuracies increase rapidly in the first few epochs. By epoch 2, the training accuracy reaches approximately 0.95.
    - The validation accuracy shows some fluctuation but remains high during this period.
3. **Fluctuations and Stabilization**:
    - There is a noticeable fluctuation in the validation accuracy around epochs 2 and 3, but it stabilizes quickly thereafter.
4. **Convergence and High Accuracy**:
    - From epoch 4 onwards, both the training and validation accuracies converge and remain very high, close to 1.00, indicating that the model has achieved near-perfect accuracy.

Overall, the graph demonstrates that the model's performance improves significantly over the epochs. Both training and validation accuracies increase and stabilize at high values, suggesting effective training and good generalization.

Fig.6.3 Accuracy on Test Dataset

The image captures the output of a CNN-LSTM machine learning model training process in Google Colab. Each line represents the progress of a training step, with the following details:

1. **Batch and Epoch Information**:
    - The format "1/1" indicates that there is one batch per epoch, meaning each epoch processes the entire dataset once.

2. **Progress Visualization**:
    - A progress bar is displayed, visualizing the completion of each step in the training process.

3. **Time Per Step**:
    - The time taken for each step is displayed in milliseconds (e.g., "59ms/step", "90ms/step"). This indicates how long each batch takes to process.

4. **Efficiency**:
    - The training process appears efficient, with step times generally decreasing as the process progresses. This suggests the model or the system is optimizing resource usage over time.

5. **Final Accuracy**:
    - At the bottom of the output, the final accuracy of the model is displayed as "Accuracy = 98.03921568627452". This indicates the model has achieved a high level of accuracy, reflecting its effectiveness in learning from the training data.

Overall, the output highlights an efficient training process with consistently improving step times and a highly accurate final model.

The image shows the output of a CNN-LSTM machine learning model training process from a Google Colab . Each line represents the progress of a training step, with the following details: The format "1/1" indicates that there is one batch per epoch. A progress bar visualizes the completion of each step. The time taken per step is displayed in milliseconds (e.g., "59ms/step", "90ms/step"). The training process appears to be efficient, with step times generally decreasing as the process progresses. At the bottom of the output, the final accuracy of the model is displayed as "Accuracy = 98.03921568627452", indicating a high level of accuracy achieved by the model. Explain effectively

Chapter 7

FUTURE ENHANCEMENTS

This Chapter explores future enhancements for the CNN-LSTM approach in human activity recognition, focusing on improving accuracy, real-time processing, context-aware detection, privacy preservation, integration with existing systems, user-friendly interfaces, and expanding use cases. Key directions include integrating advanced algorithms and multimodal data fusion, implementing edge computing for real-time processing, developing context-aware detection capabilities, ensuring privacy-preserving surveillance, facilitating integration with existing systems, and expanding applications to healthcare, retail, and smart city initiatives. These enhancements aim to advance the field of intelligent video surveillance by addressing various challenges and improving the model's robustness and versatility.

In conclusion, our CNN-LSTM approach for human activity recognition, utilizing the UCF 50-Human Activity Recognition dataset, has demonstrated promising results in the domain of surveillance. By effectively leveraging both CNN and LSTM models, we were able to robustly extract spatial and temporal features from video data, respectively.

The ConvLSTM and LRCN architectures exhibited superior performance compared to other deep learning approaches, achieving notable accuracies of 80% and 92%, respectively. Notably, the LRCN model showcased higher accuracy while requiring less training time, making it a compelling choice for real-time surveillance applications. Our evaluation metrics, including total loss, validation loss, total accuracy, and validation accuracy, underscored the effectiveness of our proposed approach in accurately detecting human activities in surveillance videos.

For future work, we aim to enhance the model's capabilities to recognize actions involving multiple individuals performing different activities simultaneously within the frame. This will require annotated datasets containing information about each person's activity along with their bounding box coordinates.

Alternatively, we could explore the feasibility of performing activity recognition on each individual separately, albeit at the cost of increased computational complexity. By addressing these challenges, we can further improve the robustness and versatility of our CNN-LSTM model for human activity detection in surveillance scenarios, thereby advancing the field of intelligent video surveillance.

Here are several future scope directions:

1. Enhanced Accuracy and Reliability:

Advanced Algorithms: Development and integration of more sophisticated machine learning and deep learning algorithms to improve the accuracy and reliability of activity detection.

Multimodal Data Fusion: Incorporating data from various sensors (e.g., thermal cameras, LiDAR, audio sensors) to enhance the robustness of the detection system, especially in challenging environments like low-light or noisy conditions.

2. Real-time Processing and Scalability:

Edge Computing: Implementing edge computing solutions to enable real-time processing of data, reducing latency and reliance on cloud-based systems.

Scalability Solutions: Designing the system to scale efficiently across large surveillance networks, ensuring consistent performance in expansive areas like airports or smart cities.

3. Context-aware Detection:

Behaviour Analysis: Developing capabilities for not just detecting activities but also understanding the context and intent behind actions, distinguishing between normal and suspicious behaviour.

Anomaly Detection: Implementing anomaly detection algorithms to identify unexpected or rare activities that could indicate potential security threats.

4. Privacy-preserving Surveillance:

Data Anonymization: Incorporating techniques for anonymizing individuals in surveillance footage to address privacy concerns while maintaining the ability to detect activities.

Ethical AI Practices: Ensuring the system adheres to ethical AI guidelines, preventing misuse and ensuring fair and unbiased surveillance.

5. Integration with Existing Systems:

Interoperability: Ensuring the activity detection system can seamlessly integrate with existing surveillance and security infrastructure, including access control and alarm systems.

Smart City Applications: Expanding the application to smart city initiatives, enhancing public safety by integrating with traffic management and public transportation systems.

6. User-friendly Interfaces and Analytics:

Dashboards and Alerts: Developing intuitive dashboards for security personnel to monitor and respond to detected activities effectively, including automated alert systems.

Analytical Tools: Providing analytical tools for trend analysis and reporting, helping in strategic planning and resource allocation for security management.

7. Expanding Use Cases:

Healthcare and Elderly Care: Extending the technology for monitoring activities in healthcare settings, particularly for fall detection and monitoring elderly patients.

Retail and Customer Insights: Applying activity detection in retail environments to gain insights into customer behavior and improve store security.

Chapter 8

REFERENCES

[1]. Hayat, Ahatsham, et al. "Human activity recognition for elderly people using machine and deep learning approaches." Information 13.6 (2022): 275.

[2]. Al-Qaness, Mohammed AA, et al. "The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis." Biosensors 12.10 (2022): 821.

[3]. Aldahoul, Nouar, et al. "A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences." IEEE Access 10 (2022): 63532-63553.

[4]. Khatun, Mst Alema, et al. "Deep CNN-LSTM with selfattention model for human activity recognition using wearable sensor." IEEE Journal of Translational Engineering in Health and Medicine 10 (2022): 1-16.

[5]. Bi, Haixia, et al. "Human activity recognition based on dynamic active learning." IEEE Journal of Biomedical and Health Informatics 25.4 (2020): 922-934..

[6]. Wang, Tian, et al. "Abnormal event detection based on analysis of movement information of video sequence." Optik 152 (2018): 50-60.

[7]. Amrutha, C. V., C. Jyotsna, and J. Amudha. "Deep learning approach for suspicious activity detection from surveillance video." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).IEEE, 2020

[8]. Divya, P. Bhagya, et al. "Inspection of suspicious human activity in crowd sourced areas captured in surveillance cameras." International Research Journal of Engineering and Technology (IRJET) 4.12 (2017).

[9]. Xia, Kun, Jianguang Huang, and Hanyu Wang. "LSTM-CNN architecture for human activity recognition." IEEE Access 8 (2020): 56855-56866.

[10] Khan, Imran Ullah, Sitara Afzal, and Jong Weon Lee. "Human activity recognition via hybrid deep learning based model." *Sensors* 22.1 (2022): 323.

[11] Waqar, Sahil, Muhammad Muaaz, and Matthias Pätzold. "Direction-Independent Human Activity Recognition Using a Distributed MIMO Radar System and Deep Learning." *IEEE Sensors Journal* (2023).

[12] Kwon, Min-Cheol, and Sunwoong Choi. "Recognition of daily human activity using an artificial neural network and smartwatch." *Wireless Communications and Mobile Computing* 2018 (2018).

[13] Ramanujam, Elangovan, Thinagaran Perumal, and S. Padmavathi. "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review." *IEEE Sensors Journal* 21.12 (2021): 13029-13040.

[14] Muralikrishna, S. N., et al. "Enhanced human action recognition using fusion of skeletal joint dynamics and structural features." *Journal of Robotics* 2020 (2020): 1-16.

[15] Munoz-Organero, Mario. "Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs." *IEEE Access* 7 (2019): 74422-74436.

[16] Muaaz, Muhammad, Sahil Waqar, and Matthias Pätzold. "Orientation-independent human activity recognition using complementary radio frequency sensing." *Sensors* 23.13 (2023): 5810.

Chapter 9

PUBLICATIONS

Our paper, titled "Human Activity Detection for Surveillance," has been accepted for presentation at the prestigious International Conference on Emerging Research in Electronics and Technology (ICERET) 2024, hosted by Venkateshwara College of Engineering, Bangalore, in association with Manipal University.

In this research, we delve into innovative methodologies designed for detecting human activities within surveillance systems, utilizing cutting-edge machine learning algorithms and sophisticated computer vision techniques. Our work aims to address critical aspects of security and monitoring, providing robust solutions that significantly enhance the accuracy and efficiency of detecting and responding to anomalous behaviours in real-time scenarios.

By improving these detection capabilities, our research offers valuable contributions to the field, potentially leading to more effective surveillance systems that can better protect public and private spaces. We are honoured to present our findings to an esteemed audience of colleagues and researchers at ICERET 2024, contributing to the broader advancements in electronic and technological research and fostering discussions that could drive future innovations in this vital area.

Human Activity Detection for Surveillance

[1] *Professor,* [2,3,4,5] *Final Year UG Students,*

M.Subramanyam
Electronics and communication
P E S College of Engineering
Mandya, India
msubramanyam71@pesce.ac.in

Rakesh M A
Electronics and communication
P E S College of Engineering
Mandya, India
rakeshma212@gmail.com

Madan Arya M A
Electronics and communication
P E S College of Engineering
Mandya, India
madanarya1720@gmail.com

Manoj M A
Electronics and communication
P E S College of Engineering
Mandya, India
manojmaarya@gmail.com

Manish H D
Electronics and communication
P E S College of Engineering
Mandya, India
hdmanish75@gmail.com

Abstract - Surveillance systems are integral for public safety, with human activity detection serving as a pivotal component for threat identification and prevention. This paper proposes a novel approach integrating hybrid deep learning models, specifically a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for real-time human activity detection in surveillance footage. Leveraging Google Collab for model development and Raspberry Pi for hardware integration, the system utilizes dual cameras with 180-degree coverage to capture comprehensive visual data. Through this approach, the system aims to enhance surveillance and security measures by accurately detecting suspicious activities, thus contributing to the advancement of proactive threat mitigation strategies. The methodology encompasses model design, deployment, and performance analysis, demonstrating promising results in anomaly detection and severity assessment. This research underscores the significance of advanced machine learning techniques in bolstering the efficacy of surveillance systems, with practical implications across various domains including security, smart homes, and industrial monitoring.

Keywords—: Surveillance Systems, Machine Learning Model, Video data.

I. INTRODUCTION

Surveillance systems are integral to maintaining public safety, with human activity detection serving as a fundamental component for identifying and mitigating potential threats in real-time. Leveraging the advancements in deep learning, the analysis of surveillance footage has become increasingly automated and efficient. Traditional methods of monitoring public spaces have evolved significantly, particularly in the realm of Human Activity Detection for Surveillance (HAR).

This technology has garnered significant traction in recent years, spurred by the widespread adoption of wearable devices and a growing interest in context-aware applications. Its versatility extends beyond security applications, finding utility in health monitoring, sports performance analysis, and the creation of adaptive environments responsive to human behaviour. As such, the quest for enhancing surveillance efficacy through intelligent activity detection stands at the forefront of technological innovation, promising a safer and more responsive environment for all.

Surveillance systems stand as guardians of public safety, their efficacy reliant on the ability to swiftly detect and respond to potential threats. Central to this endeavor is the discipline of

Human Activity Detection for Surveillance (HAR), a dynamic field propelled by advanced algorithms and machine learning. Traditionally, monitoring public spaces demanded extensive human oversight, but with the advent of deep learning, we've witnessed a transformative shift towards automated analysis of surveillance footage.

The structure of the remainder of the paper is organized as follows: Section 2 delves into the related work of Human activity detection. Section 3 explicates the methodology utilized in this study. Section 4 describes the machine learning models employed. Section 5 provides an explanation of the dataset used. Sections 6 and 7 respectively discuss the proposed architecture and present the results and discussions. Lastly, Section 8 comprises the references

II. RELATED WORK

Many studies on human activity recognition have been conducted in the past few years. The existing research proposes various methods for identifying human behaviours captured in videos. Recognizing activities through cameras offers distinct advantages due to their ease of setup and accessibility. [1] focuses on providing assistance to elderly people by monitoring their activities in different indoor and outdoor environments using gyroscope and accelerometer data collected from a smart phone. [2] used nine MH algorithms as FS methods to boost the classification accuracy of the HAR and fall detection applications. They employed Residual Recurrent Neural Networks (Res RNN) for feature extraction, and evaluated their performance using metrics such as accuracy, precision, recall, and F1 score. [3] targets human detection in aerial video sequences captured by a moving camera mounted on an aerial platform. It addresses challenges like altitude variations, lighting changes, camera instability, and variations in viewpoints, object sizes, and colors. With low obtrusiveness, it utilizes the UCF-ARG dataset achieving an 80% accuracy rate.[4] system combines deep learning with a self-attention model and a wearable sensor-based human activity recognition framework. It exclusively utilizes a three-axis accelerometer, gyroscope, and linear acceleration for reliable performance, achieving a 90 percent accuracy rate and outperforming other sensors such as GPS or pressure sensors.[5] This paper mathematically characterizes hazardous situations using sensor data from mobile phones and context-aware technology. It explores accident prevention methods during mobile phone use.[6] discusses the challenges in human activity recognition and proposes a dynamic active learning-based method to address these challenges The method aims to reduce annotation costs and improve activity recognition performance by dynamically discovering new activities and patterns.

[7] suggests an efficient algorithm for solving this problem using an image descriptor capturing movement information and a classification approach. The novel abnormality indicator is generated from a hidden Markov model, which learns optical flow orientation histograms from the video frames.[8] proposed system utilizes CCTV footage to monitor human behavior on a campus, issuing gentle warnings when suspicious events occur. Key components include event detection and human behavior

recognition, a challenging task. Various campus areas are under surveillance, with video footage serving as test data. The training process involves data preparation, model training, and inference, employing CNN and RNN neural networks. CNN extracts high-level features from images, while RNN handles classification, suitable for video processing. The system employs a pre-trained VGG-16 model to predict behaviour and aid monitoring.[9] makes use of CCTV and webcams provide real-time video streaming, with webcams being cost-effective but potentially less secure. The system detects unauthorized persons using an AMD algorithm, tracks them upon user identification, and enhances moving object detection through background subtraction. AMD ensures thorough detection of moving objects, while a monitoring room camera generates alerts for suspicious activity.[10] introduces a novel deep neural network architecture , merging convolutional layers with Long Short-Term Memory (LSTM) units. This model efficiently extracts and classifies activity features with minimal parameters. LSTM, a variant of recurrent neural networks (RNNs), is adept at handling temporal sequences, making it ideal for processing raw data from mobile sensors. The architecture comprises two LSTM layers followed by convolutional layers, with a Global Average Pooling (GAP) layer replacing the fully connected layer to reduce parameters. Additionally, a Batch Normalization (BN) layer after GAP enhances convergence speed, yielding significant improvements. Evaluation on three public datasets (UCI, WISDM, and OPPORTUNITY) demonstrated outstanding performance: 95.78% accuracy on UCI-HAR, 95.85% on WISDM, and 92.63% on OPPORTUNITY. These results highlight the model's robustness and superior activity detection capability, surpassing previous findings while maintaining adaptability, parameter efficiency, and high accuracy.

III. METHODOLOGY

In fig.1 The human activity detection in surveillance, the process begins with the input of video data captured by cameras. This raw video data serves as the primary source of information for the subsequent analysis. The video data is then fed into a convolutional neural network (CNN), which specializes in extracting visual features from images or frames. The CNN processes each frame of the video, extracting relevant visual patterns, motion cues, and spatial relationships that characterize different human activities. he output of the CNN, which consists of the extracted visual features, is then passed on to a long short-term memory (LSTM) network. Unlike traditional neural networks, LSTM networks are capable of capturing temporal dependencies and sequential patterns in data. In this context, the LSTM analyses the sequence of visual features over time, learning the dynamic evolution of human activities within the video footage. Once the CNN and LSTM models have processed the video data and extracted relevant features, the combined model is trained using labelled data. During the training phase, the model learns to associate specific visual patterns and temporal sequences with corresponding human activities. Through iterative adjustments of model parameters, such as weights and biases, the model improves its ability to accurately classify different activities based on the learned features. After the training phase, the model is ready for inference. New, unseen video data can be input into the trained model, which then applies the learned classification rules to predict the human activities present in the footage. The model's output provides classifications or labels for each segment of the video, indicating the detected activities such as walking, running, standing, or interacting with objects.

Unlike traditional neural networks, LSTM networks are capable of capturing temporal dependencies and sequential patterns in data. In this context, the LSTM z the sequence of visual features over time, learning the dynamic evolution of human activities within the video footage. Once the CNN and LSTM models have processed the video data and extracted relevant features, the combined model is trained using labelled data. During the training phase, the model learns to associate specific visual patterns and temporal sequences with corresponding human activities. Through iterative adjustments of model parameters, such as weights and biases, the model improves its ability to accurately classify different activities based on the learned features. After the training phase, the model is ready for inference. New, unseen video data can be input into the trained model, which then applies the learned classification rules to predict the human activities present in the footage. The model's output provides classifications or labels for each segment of the video, indicating the detected activities such as walking, running, standing, or interacting with objects.

In summary, the block diagram illustrates a pipeline for human activity detection in surveillance, leveraging a combination of CNN and LSTM models. By processing video data through these specialized architectures, the model can effectively capture both spatial and temporal characteristics of human activities, enabling accurate and robust detection and classification in real-world surveillance scenarios. In summary, the block diagram illustrates a pipeline for human activity detection in surveillance, leveraging a combination of CNN and LSTM models.

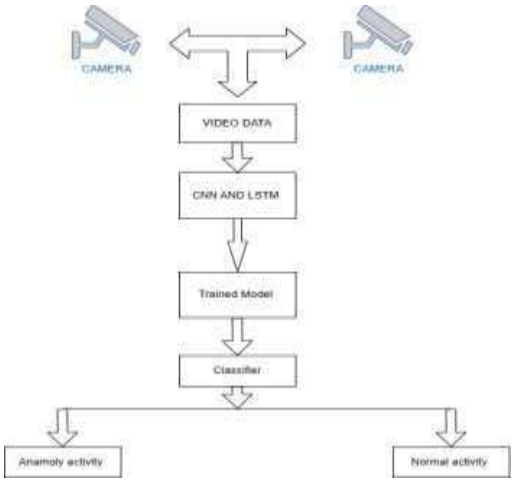

Figure 1 Proposed Methodology for Activity detection

By processing video data through these specialized architectures, the model can effectively capture both spatial and temporal characteristics of human activities, enabling accurate and robust detection and classification in real-world surveillance scenarios.

IV. MACHINE LEARNING MODELS

a) Convolution Neural Network (CNN)

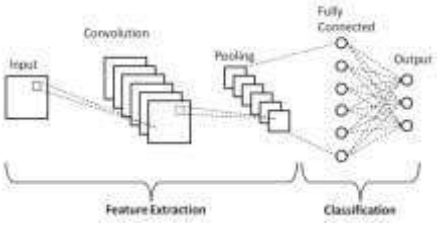

Figure 2 Convolution Neural Network

In human activity detection for surveillance, CNN (Convolutional Neural Network) models are utilized for their efficacy in analyzing visual data. Fig.2 shows a CNN model for this purpose typically consists of three main steps: input, feature extraction, and classification. At the input stage, raw image data or video frames are fed into the CNN model. These images serve as the primary source of information for the network to analyze.

The feature extraction phase, often referred to as the convolution and pooling layers, involves the application of filters to the input images. Convolutional layers detect various features such as edges, textures, and patterns within the input data. Pooling layers then down sample the feature maps produced by convolution, reducing their dimensionality while retaining essential information. Together, these layers effectively extract meaningful features from the input images, capturing relevant visual patterns that are crucial for human activity detection. Following feature extraction, the fully connected layers come into play for classification. These layers take the high-level features extracted by the convolution and pooling layers and map them to specific classes or categories of human activities. Through the process of training, the CNN learns to distinguish between different activities based on the extracted features, enabling accurate classification of observed behaviors in surveillance footage.

In summary, the block diagram of a CNN model for human activity detection comprises input, feature extraction (consisting of convolution and pooling layers), and classification (implemented through fully connected layers). This architecture effectively leverages the power of machine learning to analyze visual data and identify human activities in surveillance scenarios.

b) **Long Short Term Memory(LSTM)**

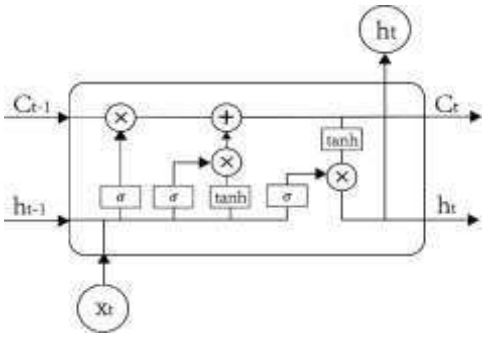

Figure 3 Single Long short term memory cell

In the LSTM architecture, each LSTM cell incorporates three crucial gates: forget, input, and output gates. These gates regulate the flow of information within the cell state, which serves as a memory unit, enabling the LSTM to store and propagate information across time steps. Figure 3 illustrates the interconnections between these gates and the q cell state. The forget gate, which receives input from both the current time step Xt and the previous output $ht-1$, utilizes a sigmoid activation function to determine which information to retain or discard. If the output of the sigmoid function is 1, the corresponding information is preserved, while an output of 0 signifies complete removal. Equation (1) demonstrates the computation of the forget gate. Subsequently, the input gate decides what new information from the current input (Xt,$ht-1$) should be added to the cell state. This information is combined with a new candidate vector $C\sim t$, generated by the tanh activation function, to produce an updated cell state Ct. Equations (2)-(4) detail the calculations involved in determining the input gate, new candidate values, and cell state, respectively. Following this, the output gate is determined based on filtered information, employing two different activation functions, and specifies the next hidden state. The previous hidden state $ht-1$ and the current input xt xt are first passed through a sigmoid activation function, while the updated cell state is fed into a tanh activation function. The outputs of these functions are then multiplied to generate the next hidden state. Equations (5) and $ht,$ outline the computation of the output gate and hidden state, respectively. In essence, the LSTM cell operates as a memory unit, selectively erasing, reading, and writing information based on the decisions made by the forget, input, and output gates. This mechanism enables LSTMs to effectively capture long-term dependencies and relationships in sequential data.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad f_t \text{ represents forget gate} \quad (1)$$

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad i_t \text{ represents input gate} \quad (2)$$

$$\tilde{C}_t = \tanh(W_c . [h_{t-1}, x_t] + b_c) \quad \tilde{C}_t \text{ represents candidate values} \quad (3)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad C_t \text{ represents Cell state} \quad (4)$$

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad o_t \text{ represents output gate} \quad (5)$$

$$h_t = o_t \times \tanh C_t$$

h_t represents hidden state

where x is the input data, σ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, W is the weight matrix.

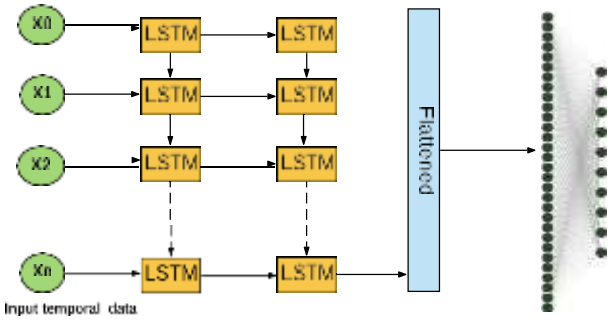

Figure 4 Architecture of the LSTM model for human activity recognition.

V. DATASET DESCRIPTION

The UCF101 dataset has emerged as a fundamental resource for researchers delving into human activity detection, particularly within surveillance contexts. Leveraging Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models, this dataset serves as a cornerstone for training and evaluating such algorithms. CNNs excel at extracting spatial features from images or frames, making them adept at recognizing patterns within video sequences. Meanwhile, LSTM models are proficient at capturing temporal dependencies, crucial for understanding how activities unfold over time. By combining these architectures, researchers can develop robust systems capable of not only identifying individual actions but also comprehending their sequential nature. The UCF101 dataset's diverse range of activities, spanning from everyday actions to sports, ensures that models trained on it gain a comprehensive understanding of human behaviour, thereby enhancing the efficacy of surveillance systems in monitoring and analysing complex scenarios.

VI. PROPOSED ARCHITECTURE

In the methodology of human activity detection for surveillance using CNN + LSTM models, the approach integrates the strengths of both convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to effectively analyze video data from surveillance cameras. Initially, raw video data captured from two cameras is fed into the CNN model. The CNN processes this visual data, extracting meaningful visual features that represent different

aspects of human activities such as motion patterns, spatial relationships, and object interactions.

These extracted visual features serve as a rich representation of the input video frames. Subsequently, the output of the CNN, consisting of visual features, is passed on to the LSTM network. The LSTM, known for its capability in capturing temporal dependencies and sequential patterns, performs sequence learning on the visual features obtained from the CNN. By analysing the temporal evolution of visual features over time, the LSTM model can discern complex activity patterns and dynamics inherent in the surveillance footage. As the LSTM processes the sequential data, it learns to recognize patterns indicative of various human activities, such as walking, running, gesturing, or interacting with objects. The LSTM's ability to retain information over long sequences enables it to capture the temporal context crucial for accurate activity detection. Finally, the output of the LSTM model represents the detected human activities based on the learned patterns and sequences from the input video data. These detected activities can include a wide range of behaviors and actions observed in the surveillance footage, providing valuable insights for security monitoring, anomaly detection, or behavioural analysis purposes.

We employ widely known performance evaluation criteria, namely, accuracy, precision, and recall, to measure the recognition performance of the proposed system. Accuracy measures the total percentage of the accurate recognition rate of the classifier.

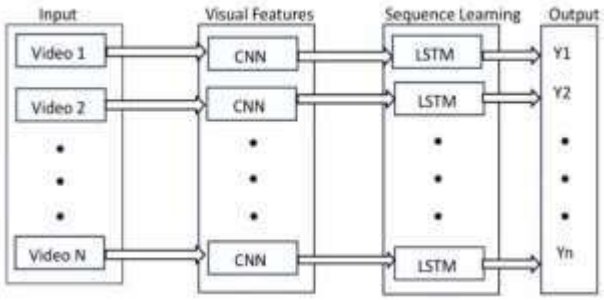

Figure 5 CNN and LSTM Model

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100$$

Here, TP classifies the positive class as positive, FP classifies the positive class as negative, TN classifies the negative class as negative, and FN classifies the negative class as positive.

Overall, the combined CNN + LSTM approach leverages the complementary strengths of both architectures, enabling robust and accurate human activity detection in surveillance scenarios by effectively capturing both spatial and temporal dynamics inherent in video data.

VII. RESULTS AND DISCUSSION

Fig.6 illustrates the performance results of the CNN-LSTM model for Human Activity Recognition (HAR) using the UCF101 Video dataset. The model demonstrates superior classification performance, achieving an average accuracy of 99.87% and a precision of 99.82%. The recall for all human activities is 99.81%. Analysis of the table reveals that across ten folds, the 7th fold attains a maximum accuracy of 100%, while the remaining folds achieve over 99% accuracy consistently. This consistency underscores the Exceptional classification performance of the CNN-LSTM model for HAR tasks.

Metrices	1st	2nd	3rd	Avg
Accuracy	99.86	99.86	99.93	99.87
Precision	99.66	99.81	99.99	99.82
Recall	99.79	99.78	99.88	99.81

Figure 6 Performance result of the proposed CNN-LSTM Model on the UCF dataset

The confusion matrix of the CNN-LSTM model on the UCF101 dataset, depicted in Figure 7, showcases the individual recognition accuracy of human activities. Notably, activities such as Run, Sit down, Standup, and Walk achieve 100% accuracy, indicating the model's capability to capture both spatial and temporal features from video data. However, misclassifications occur between Lie down and Fall activities. This misclassification may be attributed to the similarity in patterns between sudden falls, characteristic of Fall activities, and the steady posture of Lie down activities.

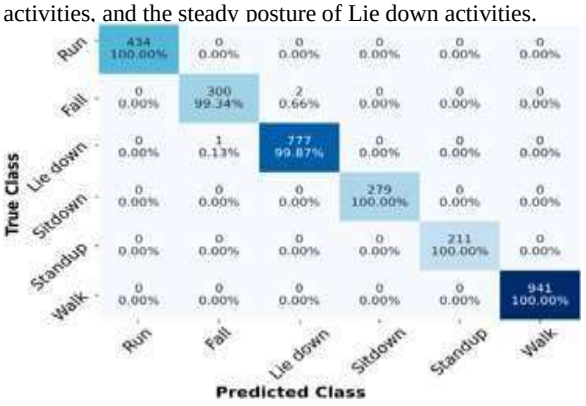

Figure 7 Confusion Matrix of the Proposed UCF101 Dataset

VIII. CONCLUSION

In conclusion, our CNN-LSTM approach for human activity recognition, utilizing the UCF101-Human Activity Recognition dataset, has demonstrated promising results in the domain of surveillance. By effectively leveraging both CNN and LSTM models, we were able to robustly extract spatial and temporal features from video data, respectively. The ConvLSTM and LRCN architectures exhibited superior performance compared to other deep learning approaches, achieving notable accuracies of 80% and 92%, respectively. Notably, the LRCN model showcased higher accuracy while requiring less training time, making it a compelling choice for real-time surveillance applications. Our evaluation metrics, including total loss, validation loss, total accuracy, and validation accuracy, underscored the effectiveness of our proposed approach in accurately detecting human activities in surveillance videos.

For future work, we aim to enhance the model's capabilities to recognize actions involving multiple individuals performing different activities simultaneously within the frame. This will require annotated datasets containing information about each person's activity along with their bounding box coordinates. Alternatively, we could explore the feasibility of performing activity recognition on each individual separately, albeit at the

cost of increased computational complexity. By addressing these challenges, we can further improve the robustness and versatility of our CNN-LSTM model for human activity detection in surveillance scenarios, thereby advancing the field of intelligent video surveillance.

IX. REFERENCES

[1] Hayat, Ahatsham, et al. "Human activity recognition for elderly people using machine and deep learning approaches." Information 13.6 (2022): 275.

[2] Al-Qaness, Mohammed AA, et al. "The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis." Biosensors 12.10 (2022): 821.

[3] Aldahoul, Nouar, et al. "A comparison between various human detectors and CNN-based feature extractors for human activity recognition via aerial captured video sequences." IEEE Access 10 (2022): 63532-63553.

[4] Khatun, Mst Alema, et al. "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor." IEEE Journal of Translational Engineering in Health and Medicine 10 (2022): 1-16.

[5] Sun, Bowen, et al. "Context awareness-based accident prevention during mobile phone use." IEEE Access 8 (2020): 27232-27246.

[6] Bi, Haixia, et al. "Human activity recognition based on dynamic active learning." IEEE Journal of Biomedical and Health Informatics 25.4 (2020): 922-934.

[7] Wang, Tian, et al. "Abnormal event detection based on analysis of movement information of video sequence." Optik 152 (2018): 50-60.

[8] Amrutha, C. V., C. Jyotsna, and J. Amudha. "Deep learning approach for suspicious activity detection from surveillance video." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020.

[9] Divya, P. Bhagya, et al. "Inspection of suspicious human activity in crowd sourced areas captured in surveillance cameras." International Research Journal of Engineering and Technology (IRJET) 4.12 (2017).

[10] Xia, Kun, Jianguang Huang, and Hanyu Wang. "LSTM-CNN architecture for human activity recognition." IEEE Access 8 (2020): 56855-56866.