Revised report

by Bharti Chugh

Submission date: 22-May-2025 08:14PM (UTC+0530)
Submission ID: 2682248181
File name: Revised_Report_1.docx (1.85M)
Word count: 10023
Character count: 59218

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

A

Project Report

on

SUSPICIOUS ACTIVITY RECOGNITION FROM LIVE VIDEO USING DEEP LEARNING

submitted as partial fulfilment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

Computer Science and Engineering

by

Rishabh Kumar Panthri

(2100290100133, CSE)

Mohan Paliwal (2100290100098, CSE)

Under the supervision of

Ms. Nishu Gupta (CSE)

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Date:

Signature

Name: Rishabh Kumar Panthri

Roll No.: 2100290100133

Date:

Signature

Name: Mohan Paliwal

Roll No.: 2100290100098

CERTIFICATE

This is to certify that the Project Report entitled "Suspicious Activity Recognition from Live Video Using Deep Learning" submitted by Rishabh Kumar Panthri and Mohan Paliwal in partial fulfillment of the requirements for the award of the degree of B. Tech. in the Department of Computer Science & Engineering and Department of Information Technology of Dr. A.P.J. Abdul Kalam Technical University, Lucknow, is a record of the candidates' own work carried out under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Ms. Nishu Gupta
(Project Guide)
(Department of CSE)

Dr. Vineet Kumar
Sharma
(Dean CSE)

ACKNOWLEDGEMENTS

We express our sincere gratitude to our supervisor Nishu Gupta, Department of CSE, KIET, Ghaziabad for their invaluable guidance, continuous support, and encouragement throughout the project. Their dedication and technical expertise have been a source of inspiration. We also extend our gratitude to Dr. Vineet Kumar Sharma, Dean(CSE) for his support and assistance during our work. Lastly, we acknowledge our friends and faculty members for their help and motivation.

Date:

Signature

Name: Rishabh Kumar Panthri

Roll No.: 2100290100133

Date:

Signature

Name: Mohan Paliwal

Roll No.:2100290100098

ABSTRACT

Deep learning techniques have largely improved the efficiency of surveillance systems that can now monitor a space for unusual or abnormal behaviour in real-time with an increased rate of accuracy. The present work is dedicated to the creation and application of the intelligent system capable of recognizing activities by means of video footages, and the UCF101 dataset has been chosen for this end purpose, as it covers a broad spectrum of human moves. The actions of people on the video clips in this database can be seen from many different points of view, plus the people can be in lots of different positions, and the lighting can be very different indeed. The system proposed here is divided into a few stages of processing that implement spatial and temporal analysis for efficient activity recognition. The system's first stage entails the identification of all the main objects in every video frame with the help of the YOLO (You Only Look Once) object detection model. The speed and accuracy of YOLO in detecting and locating individuals or objects across the video frames are taken into account as significant features of this model. Once the subjects are identified, they are concentrated and then fed into the ResNet-based convolutional neural network, which is used to categorize the regions spatially. The historical dynamics of the actions in the videos are represented by the used of a Transformer-based model. In this context, the model processes the sequence of frame-level features extracted from the ResNet model, and it learns the temporal patterns of these features to correspond to certain events. The Transformer defines the action in a more general way not only from the primitive motion descriptions but also in terms of temporal information including the change from one action to the next. In other words, different actions can be very similar to each other, but they can be different only in their motion.

An extraordinary aspect of this system is that it can work instantly, i.e., it is capable of operating in live surveillance environments without any problems. If the system recognizes pre-defined suspicious activities, which consist of aggressive movements or unauthorized access, it will immediately trigger an alarm. As a result, law enforcement officials will be able to take timely actions, thereby possibly stopping the incidents before they get worse. This system brings together the best elements of YOLO, ResNet, and Transformer architectures and provides a solid, interpretable, and computationally affordable way of carrying out event recognition.

TABLE OF CONTENTS Page No.

LIST OF FIGURES AND TABLES Page No.

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

Lately, the deep learning revolution has been a game-changer for many sectors, and video surveillance is one of them. Conventional surveillance systems majorly depended on human supervision and that was not only time-consuming but also it could lead to errors and inattention quite easily.

Since the number of surveillance data has exploded along with the extensive distribution of cameras not only in public places but also in private ones, manual monitoring has significantly shown its limits.

Surveillance systems based on deep learning have brought a game-changing option through quick, automatic video content recognition. At the core of these systems, the latest breakthrough neural networks are responsible for handling vast amounts of imagery and detecting abnormal and/or specific cases that are nearly invisible to human eyes. The effect of this transformation is that a passive surveillance model shifts to an active mode where the security infrastructure can respond, perform, and learn.

Through the adoption of the technologies of object tracking, detection, classification, and behaviour recognition through time, the new surveillance systems automatically support themselves to high-level competence by recognizing the most intricate activities with a mere slip of accuracy which consequently brings a new level of awareness and safety altogether.

1.2 PROBLEM STATEMENT

Despite the progress in video analytics, most surveillance networks still encounter major problems that hinder their effectiveness in real-world deployments. The absence of real-time processing capabilities combined with high accuracy is still the biggest issue, as such capabilities are important for security systems that are time-critical.

Almost every traditional system is either spatial or temporal, which are not seamless for activity recognition due to their inappropriate integration of both aspects.
Moreover, the interoperability with environmental changes like different lighting conditions, camera views, and crowded scenes is one of the issues. Since the system cannot effectively commit to a particular mode of operation, its deployment in different environments becomes a huge challenge.

This endeavour introduces a unified approach that exploits new generation deep learning models powered by the combination of YOLO (object detection), ResNet (frame-level classification), and Transformers (temporal analysis) to tackle these challenges. The goal is not only to produce high-quality activity detection but also to achieve it in real-time with few or no delays, thus enabling it to be deployed in the complex and harsh scenarios.

1.3 RESEARCH MOTIVATION

The uptick in demand for automated surveillance solutions can be attributed to the variety of security issues faced by the public and private sectors. Notably, the growing pervasiveness of the video surveillance solutions used in cities, shopping and business centres, transportation hubs, and residential areas has made the work of human operators more challenging. Thus, the monitoring of dozens of video feeds simultaneously by human beings is neither practicable nor avert of slow threat detection and response.

To put it another way, the cellular security field right now needs adaptable systems that offer automatic responsiveness to constantly changing situations. To illustrate, the behavior that is perfectly normal in one situation can be viewed as suspicious in other cases. It is, therefore, clear that such dynamic configurations are required to watch over the surrounding circumstances and apply pre-set rules for the detection of irregularities.

The main motive behind this research is to build a model of surveillance that is flexible and smart—named ADAG (Activity Detection and Alert Generation)—to reach the goal of an operational and effective camera system that works on its own. ADAG is intended to find context-dependent

suspicious activities and generate alarms in real-time, thus ensuring that timely responses are made and safety is maintained. By combining state-of-the-art deep learning models and an environment detection and localization, this approach intends to push the horizon of modern surveillance to a new level.

CHAPTER 2

LITERATURE REVIEW

Event-based detection of human behavior using video has in recent times been an integral part of smart surveillance systems, mainly focusing on suspicious or illegal activities that are happening in real-time. With the upsurge in security threats and also in the geometric progression of storage of visual data, a number of researchers have put their hearts into developing deep learning-based systems that carry out automated behavior recognition.

Among the initial endeavors in this domain was the one that came from Kauthkar and Pingle. They are the ones who created a detection framework which is able to execute hostile actions recognition like punching, kicking, knife attacks, and gunfire. They did not stop there, though, and even went further to incorporate an automated email alert system that was integrated into the whole setup to provide real-time information to the authorities concerned [1]. From there, Chole et al. went on to design a modular system that is made up of three main components, i.e. video capture, preprocessing, and feature extraction. They employed the KTH dataset which has a global acceptance for their experiments to build and test their architecture for simple human activity recognition [2].

The technology of Convolutions Neural Networks (CNNs) was quite instrumental in the exercise of activity detection. For instance, Devi et al. first designed a CNN model that could identify confirmed unusual behaviors in video surveillance by using the first step of the CNN-based feature extraction. They also put into practice an alert-emailing feature which is run by API for instant threat warning [3]. Furthermore, the idea was developed to the point where Kumar et al. dabbled in various deep learning models, including Hybrid LSTM, Time Distributed CNN, and Gated Recurrent Units (GRU) achieving promising accuracy and motif detection in different real-world settings [4].

It was a great success when Mohamed et al. suggested a system that was a combination of YOLOv8 for object detection, ResNet34 for deep feature extraction, and Transformer encoder layers for temporal modeling in their paper.

The coming together of these three models brought about a model that was fast enough to be in real-time deployment with temporal coherence and in turn, resulted in a new standard in the field of motion detection systems [5].

Long short-term memory (LSTM) networks have been famous for a good while in video sequence understanding, too. Amrutha C.V. et al. used LSTM as well but their activation function was ReLU and they have also added dropout layers to their model. They used 7,035 several labeled frames from more than one dataset and was also able to get good results for [6] because of the LSTM model in sequential frame analysis. The work was rechecked and results were improved further by Gowshikhaa et al., who have given more details on the feature of the time difference extraction such as the one being the most contributory to the surveillance analysis for better results [7].

It has been discovered that the models which are purely visual might fall short; thus, many experts are testing multimodal systems. Chen, for example, has combined such a system (which is as smart as a security system) with physical alarms based on motion sensors and thereby was able to get quick results [8]. At the same time, Gugale et al. skilfully performed their duties in the area by building a part of it consisting of five categories - namely, shooting, boxing, sword fighting etc. that was very useful for models of classification due to its variety [9].

The main aim of the research is to develop CNN with improved efficiency. Indhumathi and Balasubramaniam picked a Kaggle database of 2160 color images, which significantly increased the efficiency of their CNN as they utilized it achieving a high performance [10]. On the other hand, the study by Buttar et al. sought to exploit the problem faced by CNN-based detection systems, namely, the lack of trust and reliable application of the models [11]. CNNs have also been investigated by Quadri and Katakdhond for real-time crowd detection in crowded dynamic scenarios and have been found efficient and scalable in that context.

A thorough analysis of deep learning models for video surveillance has been conducted by Namithadevi et al. in their previous papers, whereby they proposed the use of a hybrid model combining LSTM and MobileNetV2 for real-time analysis, which turned out to be of the highest impact [14, 15]. In addition, the study of Singh et al. has made the improvements maximized, and

the detection has been gotten rid of in detail hence fast alterations made possible [16].

Emerging research has been focused on the integration of new architectures where the researchers just as described in the next sentence. The researchers were Khushi et al. who performed innovative experiments of combining VGG-19 and traditional CNN models as they were able to achieve better results in feature representation [17], and Pramanik created a brand new transformer-based Deep Reversal Attention Network for multi-sensory action recognition tasks [18]. Raza et al. developed crowd analysis techniques that were meant only for those situations in a crowd where they can catch the abnormal behavior [19], and Tiwari et al. were the authors of the deep learning methods used for autonomous threat identification proliferation [20].

Advanced models using machine learning techniques have replaced some of the earlier ones and this is a great example of how it works. This was suggested by Mudgal, and Punj as a means of solving the problem of action recognition attributes that are very subjective [21], whilst Ahmed decided to build a powerful pose descriptor that also improved recognition accuracy [22]. An extensive review was published by Tripathi et al. that not only counted but also described in detail the process of recognition of behavior that takes place from the very beginning until now [23], and Ayed et al have put forward the argument that if facial emotion recognition is to be accurate then emotion recognition has to have access to the full face the same way expression recognition depends on the full face [24].

Ilidrissi and Tan presented the implementation of an assembly of the three models in the section “Recently published” [25], Verma et al. dealt with the existing methods in the field by conducting a comprehensive study of the supervised, semi-supervised, and unsupervised learning techniques that are the base of the new AI systems [26], whereas Selvi et al. is a researcher that had his CNN model optimized [27]; modeling where the researchers “rebenchmarked” a pretrained CNN model, increased the number of parameters and ran it on new data.

The areas were not only kept but also significantly extended by many researchers who focused on not only behavior recognition task optimization but expanding the portions of the latter. This was well-demonstrated in the latest research papers of Genemo’s intelligent exam supervision

implementation [28] and Saba et al.'s network architecture for detecting complex activities within the range of L4 branches [29].

Barathi et al. embarked on a project that aims to develop CNN architectures for use in the surveillance sector that are more efficient especially in live streaming, selecting, and tracking of the surveillance video [30].
Together, these studies reiterate the swift and continuous improvement in the creation of smart, precise, and instantly responsive surveillance technologies. The combination of spatial, temporal, and contextual learning is still the major force for change and is making the industry approach more and more the ultimate self-sufficient and flexible security solutions.

As the research advances, it becomes more and more important to make these systems that are not only machine learning-driven but also context-aware, interpretable, and able to adapt to the never-experienced-before situations. One of the future directions might be the incorporation of the multimodal data—audio, thermal imaging, and biometric inputs—besides, if used adequately, that can serve to enhance detection robustness.

Also, significantly boosting the widespread deployment of the next generation will be the utilization of the so-called "lightweight" models that can run on edge devices. These improvements will not only increase the ability of the system to be highly reactive and precise but will also lead to the expansion of the intelligent surveillance system usage in sectors such as public safety, smart cities, industrial, and essential infrastructure security with no geographical limitations.

CHAPTER 3

PROPOSED METHODOLOGY

The described model is based on a combination of deep learning and its derivatives and is dedicated to the automated detection, and the semantic segmentation of the human activities in video feeds for the real-time surveillance system. The system by means of the YOLOv8, ResNet-34, and Transformer Encoder networks is equipped to realize a higher accuracy of the recognition of the activities with high energy efficiency, and in turn, makes it the most suitable for real-time surveillance applications. The user interface engages the in-browser inference engine of HTML5, and WebRTC, which ensures real-time detection and the video interface to be the most user-friendly. The backend, constructed with Flask and PyTorch, is responsible for all the essential tasks like model execution, frame analysis, and prediction.

3.1 SYSTEM ARCHITECTURE

The system architecture includes three leading elements that operate in complete harmony to provide a human monitoring end-to-end video-based human activity recognition framework:

1. Activity Classification Pipeline:

- The use of YOLOv8 for person detection in video streams and extraction of frames in real time. It solves areas of interest (ROIs) where people are engaged in activity.
- ResNet-34, a pre-trained deep convolutional neural network, has been retrained on the task of surveillance image classification to extract the most appropriate spatial features from each recognized frame. The model is fine-tuned for the task of surveillance-based image classification.
- The next step is using Transformer Encoder layers to model temporal correlations within the sequence of video frames. Thus, the system is not just told what is happening in a single frame but is capable of recognizing how an action evolves over time.
- The created head of the classification finalizes the classification process taking temporally aligned features into account. Thus, an event is predicted by analyzing the features that are

aligned in time.

2. Frontend System:

- With the use of HTML5 and WebRTC, a real-time video interface was created that allows the users to upload or stream surveillance videos directly in the browser.
- The frontend provides alert visualization, in case of anything that seems suspicious or predefined actions, an alarm will be shown in the form of a pop-up, visual overlays, or e-mail notifications.
- Users can perform different interactive actions such as controlling the video timeline, doing frame-by-frame annotations, and viewing predictions that run in real-time.

3. Backend Infrastructure:

- Flask and PyTorch are the technologies used in too of model training, video processing, and server-side inference.
- The system gets the cloud infrastructure ready for the sake of scalability and quickness of response, which allows the distributed processing and storage.
- Using HTTPS for all communications makes it impossible for unauthorized persons to intercept it, and adding end-to-end encryption ensures the privacy of the data and secure transfer between the clients and the server

18

3.2 DATA PREPROCESSING

It is a must-have step to come up with a drone that could be a pilot's right hand in every situation from the preparation of raw video data for the training and/or robust model inference. The process is carried out in the following way in sequence:

- Frame Extraction: The initial videos are divided into an array of images where each image is a frame at a constant frame rate, usually not less than 30 frames per second (FPS) to provide

temporal consistency.

- The initial videos are divided into an array of images where each image is a frame at a constant frame rate, usually not less than 30 frames per second (FPS) to provide temporal consistency.
- Bounding Box Detection: The model for the bounding box (YOLOv8) is applied to all frames for the detection of people. After this operation, the bounding box is cropped and the isolated regions of the human body become the next stage for the operations.
- The model for the bounding box (YOLOv8) is applied to all frames for the detection of people. After this operation, the bounding box is cropped and the isolated regions of the human body become the next stage for the operations.
- Image Resizing: The small regions of the human are enlarged so as to fill the input size of the ResNet and Transformer models, both equal to 224 × 224 pixels.
- The small regions of the human are enlarged so as to fill the input size of the ResNet and Transformer models, both equal to 224 × 224 pixels.
- Normalization: It is highly beneficial to the stability and speed of the training, if the pixel intensity is in the range of [0, 1] when values are multiplied by the reciprocal of a constant.
- It is highly beneficial to the stability and speed of the training, if the pixel intensity is in the range of [0, 1] when values are multiplied by the reciprocal of a constant.
- Data Augmentation: The only solution to extend the scope of the model's capability is to generalize the model and to reduce the overfitting by the following augmentation process such as random rotations, horizontal flipping, brightness variation, and scaling.
- To improve model generalization and prevent overfitting, augmentation techniques such as random rotations, horizontal flipping, brightness variation, and scaling are applied.
- Implementing augmentation strategies, like random rotations, horizontal flipping, brightness alterations, and scaling, helps to make the model generalize better and not overfit.
- Feature Extraction: The frames that have been preprocessed are inputted into the ResNet-34 model to obtain the feature maps for each image with a high dimensionality. The sequence of

levels of details was created temp

- Dataset Splitting: The entire dataset is split into 3 partitions, namely 80/10/10 which indicate the percentage of the dataset that goes to training, validation, and testing, respectively (training as 80%, validation as 10%, testing as 10%). The choice of such a division allows describing the former and leaving the latter percentage for preparatory activities associated with the deployment of the model.'

3.3 MODEL TRAINING

The model is trained majorly in two phases:

ResNet-34 Fine-Tuning:

A pre-trained ResNet-34 model is fine-tuned on the action recognition data. Fully connected layers at the end are changed to the number of classes for the output (UCF101 dataset) and the weights are then updated through the process of backpropagation.

A pre-trained ResNet-34 model is fine-tuned on the action recognition data. The last fully connected layers are modified to match the number of output classes (UCF101 dataset), and weights are updated using backpropagation.

Transformer Encoder Training:

The model Transformer is trained on the feature sequences that are extracted learning to model time patterns and predict activity labels on a series of frames. Positional encodings are added to help the model know the order of frames and attention mechanisms are used to

focus on significant parts of the video sequence.

The model Transformer is trained on the feature sequences that are extracted and used to model time patterns and predict activity labels on a series of frames.

Positional encodings are used to inform the model which position of the current video frame is, while attention mechanisms help the model to concentrate on critical situations in the video. Loss functions such as cross-entropy, in addition to optimizers like Adam or SGD, are applied to lower the prediction error. Training is performed by iterating the epochs along with batch-

wise processing running on GPU for hastening the convergence process.

3.4 ALERT GENERATION

After the model is finished with the suspicious or predetermined activity, the real-time alert notification system gets on the go. A system that uses JavaScript as a base and perfectly fits the front-end interface and brings with it the following features:

- Notifications that are delivered as e-mail notifications so that security officers are alerted with the cyber threats detected.
- The use of visual alarms like pop-ups to attract instant attention is provided.
- The events that are detected are being stamped with their exact time and in that case they can be used for reference in the future.

These alerts are set up based on the benchmarks that can be modified as per customizable activity thresholds and user-defined parameters, that way the system is far from a one-size-fits-all and can be custom-tailored to the specific security needs of the given environment and operation.

CHAPTER 4

PROPOSED SYSTEM

With the aim of achieving accurate and real-time human activity recognition, the proposed system is based on the use of various deep learning models, each of which is developed for a specific part of the video analysis pipeline. The system consists of three essential neural network models, YOLOv8, ResNet34, and a Transformer Encoder, that have been combined in a single framework to carry out the actual video processing work, understanding the activity in both spatial and temporal dimensions.

1. YOLOv8 – Real-Time Person Detection

YOLOv8 (You Only Look Once, version 8) is the first model that the system utilizes to accomplish the task and is there to carry out the detection of persons in the video, the real-time feature.

- Speed and Accuracy: YOLOv8 is a system that has won a reputation owing to its detection speed and precision, which are the main points for an application that is called real-time. The single-step system allows objects to be determined directly without the need for more complex region proposals, making it faster.
- Bounding Box Localization: Once a video is parted into the frames, YOLOv8 handles each frame on the base of which human subjects are localized precisely with the help of bounding boxes.
- Noise Reduction: The approach of noise reduction is to separate the regions containing people from the background, which in effect, weakens the signal considerably, and the next models can concentrate only on the carrying of the relevant features.
- Scalability: YOLOv8 is designed for the processing of both edge and server levels, which allows it to change from the lightweight version to a high-performance one easily. Edge devices are those that perform on-site and as a result, these services tend to use cheaper, or perhaps even obsolete machines; server-level here means from the supplier's end, and such a level generally implies the use of very powerful computer systems in the data store.

2. ResNet34 – Extracting Features from the Spatial

As soon as the human regions are detected and cut out from the video frames, they are passed on to the ResNet34 model, which is a 34-layer deep convolutional neural network, pre-trained on ImageNet and additionally, fine-tuned for this specific action recognition task.

- Residual Learning: ResNet architecture adopts skip connections to avert vanishing gradients as well as allows deeper networks that capture more abstract visual patterns formed.

- Visual Representation: ResNet34 pick out rich spatial features like body posture, clothing movement, and contextual cues in each of the frames, then these are further encoded into high-dimensional vectors.

3. Transformer Encoder – Temporal Modelling and Activity Recognition

The Transformer Encoder is responsible for understanding time and action patterns which is a pre-requisite for activity recognition.

- Temporal Dependency Capture: Where ResNet34 processes each frame individually, the Transformer looks at the relationship between the frames, i.e over time, so the model can detect patterns like possibly dangerous behaviour or repetitive gestures.

- Self-Attention Mechanism: Transformers use self-attention to measure the contribution of each frame in the context of the whole sequence. So the mechanism can capture long range dependencies that are not easily found in standard RNNs or CNNs.

- Contextual Awareness: By looking at frame-to-frame connections the chi-model can track the development of an action like raising a hand followed by a swing which might mean a punch even without much training.

- Scalability and Parallelism: The Transformer unlike RNNs allows parallelism in training and converges fast so it's perfect for large video datasets.

4.1 INTEGRATION AND WORKFLOW

Workflow of the proposed system as a whole is represented by the following sequence in a very simplified manner:

- Input Video → Frame Extraction
- Frame → YOLOv8 → Cropped Human Regions
- Cropped Regions → ResNet34 → Feature Vectors
- Sequence of Feature Vectors → Transformer Encoder
- Transformer Output → Classification Layer → Activity Label

4.2 KEY BENEFITS OF THE PROPOSED SYSTEM

- Real-Time Surveillance: Combination of YOLOv8 and the possibility to use the small model of TinyML with JavaScript in the web browser enables the system to achieve low-latency operation.
- Improved Accuracy: The system also removes possibly unnecessary additional classes of objects which increase misclassification errors by separating human findings and applying the two-dimensional and time-related filtering mechanisms.
- Modularity and Expandability: The different hardware parts in the system can be easily replaced or enhanced (e.g., for downward substitution of ResNet with MobileNet, or the relatively new EfficientNet, which is energy efficient for edge devices), thereby catering to the diversity of operational scenarios.
- Contextual Decision-Making: With The Transformer's, the machine can think at a level higher than the basic and thus gain a clearer insight into human behavior. It becomes quite simple for it to be able to recognize the moves as usual or suspicious.

4.3 ARCHITECTURE OF THE PROPOSED SYSTEM

Fig. 4.1. outlines the architecture and training sequence of the hybrid model:

1. Feature Extraction with Pretrained ResNet:
   The input frames (previously cropped and preprocessed) are first passed through a pretrained ResNet-34 model, which extracts high-level visual features from each image.

2. Dimensional Projection:
   The extracted features are projected to a fixed dimensional space (512 dimensions) to ensure compatibility with the subsequent Transformer layers.

3. 4-layer Transformer Encoder:
   A Transformer encoder with four layers is employed to model the temporal dependencies and sequential patterns present across video frames. This captures the contextual relationships between consecutive frames in a video sequence.

4. Temporal Attention Pooling:
   After encoding the temporal structure, a temporal attention pooling mechanism is applied to aggregate information across the sequence by assigning different weights to frames based on their importance in recognizing the activity.

5. Classification Head:
   The pooled features are passed to a classification head that maps the sequence representation to an activity class from the predefined set (e.g., from the UCF101 dataset).

6. Model Training:
   The hybrid model is trained with specific hyperparameters:

- Optimizer: AdamW
- Learning Rate: 1e-4 for ResNet and 2e-4 for the Transformer
- Learning Rate Scheduler: Cosine annealing

- Loss Function: Cross-Entropy with label smoothing
- Batch Size: 16
- Training Epochs: 43

7. Evaluation:
   Once training is completed, the model's performance is evaluated using standard metrics such as accuracy, F1-score, and confusion matrix analysis.

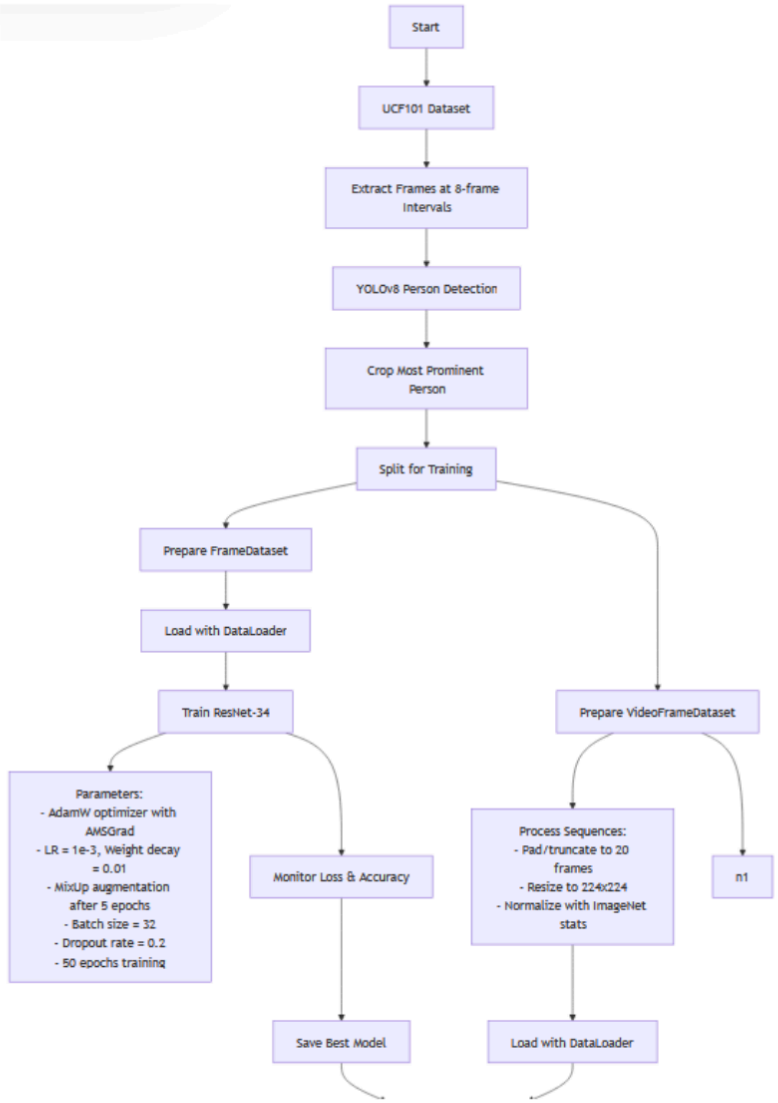

Fig. 4.1. Architecture Of Proposed System

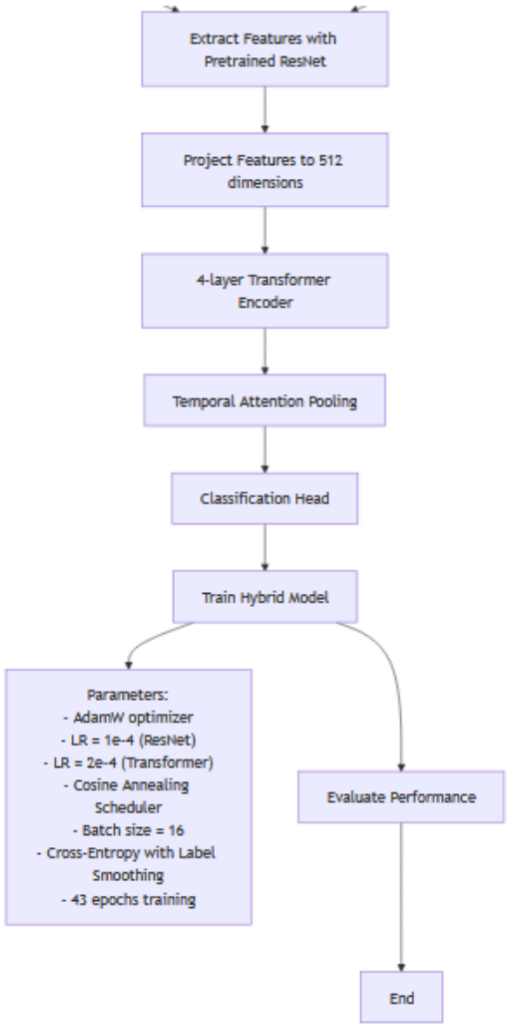

Fig. 4.2. Architecture Of Proposed System (continued from prev. page)

Dealing with the data pipeline, Fig. 4.2. is supposed to illustrate the stages of preprocessing and the primary training of the model:

- Start & Dataset Selection: The initial step in the execution of the process is the UCF101 dataset which is taken, this dataset is one of the most popular benchmarks used for human action recognition.

- Frame Extraction: One part of the video every 8 frames is taken to satisfy the requirements of both compute efficiency and temporal coverage.

- YOLOv8 for Person Detection: Your Latest One (YOLO) version 8 is a great example of an object detection model that is used for the identification of people in particular frames through positioning them.

- Cropping the Most Prominent Person: The person with the highest level of focus is the one to be chosen out of the recognized ones thus through the cropping this single person is the entity that the audience focuses on as a doer of the action.

- Split for Training: Following the processing of the data, it is divided into two paths: Initially, a branch that is intended to train a ResNet-34 model with individual frames (spatial learning). The second branch is meant to derive a video sequence dataset that will be used by the Transformer (temporal learning).

- Initially, a branch that is intended to train a ResNet-34 model with individual frames (spatial learning).

  The second branch is meant to derive a video sequence dataset that will be used by the Transformer (temporal learning).

Left Branch – Frame-Based ResNet Training:

Prepare FrameDataset:

The frames were cropped, classified, and loaded to a dataset.

Data Loading:

With PyTorch DataLoader, the dataset is loaded for efficient training.

Left Branch – Frame-Based ResNet Training:

Prepare FrameDataset:

Frames were cropped, classified and loaded into a dataset.

Right Branch – Sequence Preparation for Transformer:

Prepare VideoFrameDataset:

Frames are selected in sequences (20 per video) and data rate uniformity is maintained by padding or truncation.

Preprocessing:

Each frame is resized to 224x224 and normalized using ImageNet stats before fed into the ResNet.

Preprocessing:

The dimensions of each of the frames is resized to 224x224 and the frames are normalized using the ImageNet stats before being fed into the ResNet.

Data Loading:

Subsequently, the sequences undergo transformations, and through the use of a DataLoader, they are now ready to be fed into the Transformer Encoder.

4.5 FEATURES OF PROPOSED SYSTEM

The proposed hybrid action recognition system has used fresh deep learning methods that are available today and has deployed an arsenal of architectural innovations to help identify accuracy,

train efficiency in less time, and make the system suitable for real-time use-cases. Here are the main features and advantages of the system that are worth considering:

1. Person Scoring & Selection

Through the scoring mechanism based on bounding box metrics, the system ensures that the focus is maintained only on the most relevant person in a scene, particularly in the case of a crowded environment. The person with the most central and largest bounding box is selected for further analysis. With such scoring, the model's ability to locate the main subject of interest immediately is highly enhanced, which in turn keeps the misclassification rate to a minimum. A person's recognition factor is paramount in a surveillance situation where multiple individuals appear at the same moment, but only one is the potential threat actor.

2. Smart Frame Sampling (Interval = 8)

The process of making the video a source of frames is still very common in the field. However, every frame cannot be used effectively due to certain computationally heavy and redundant frames. The system samples every 8th frame of video. It is the stride of the frames that remain fixed as a result of the movement of the legs. In other words, it is a measure of the distance needed to move from every point in time to the next point in time. The procedure is capable of capturing crucial motion patterns and thus at the same time avoiding the same images and resulting in training speedup and less computational load. With a uniform step that separates the two sampling frames, not only the variations in movement occurring in a sequence are taken into consideration, but also the model is prevented from becoming overwhelmed by superfluous information.

3. Hybrid ResNet + Transformer Architecture (Compared to 3D-CNNs or LSTMs)

The most popular methods such as 3D CNNs and LSTMs are trying to solve this problem. For example, 3D-CNNs always use additional time to obtain results, and LSTM-based devices usually consume a long computing time.

- The use of ResNet-34 makes it easy to identify spatial features more effectively from each frame, hence, the model can be applied to capture the very details of an image.

- Transformer Encoder is a mechanism that models the sequential dependencies across the frames by the introduction of multi-head attention. Therefore, the new model elegantly solves 3D-CNNs' and LSTMs' issues with accuracy and efficiency, especially when it comes to long video sequences.

4. Strong Regularization Techniques

One of the things that we can do to minimize overfitting and to maximize model generalization is to make the following regularizations:

- MixUp Augmentation: The method randomly mixes pairs of training examples, which makes it easier for the model to learn smooth and clear decision boundaries.

- Label Smoothing: It is a technique that changes the one-hot encoded labeled data to fractional values in an attempt to dull the sharp edges of the model's confidence brought about by the labels used to train the model.

- Dropout: The approach is to randomly drop nodes in the network as the model is being trained so as to make sure no two or more neurons become co-dependent on each other.

- Gradient Clipping: It changes the gradient that is going to be back-propagated up the network in order to make sure that the gradients don't explode as the model gets deeper. In practice, this set of methods render the training process relatively more consistent and less volatile, particularly if one is working with video datasets that are of a large size.

5. Learnable Attention Pooling

Rather than using simple pooling methods (such as average or max pooling) that do not consider feature relevance and are static in nature, the model makes use of learnable attention pooling that adaptively assigns frame-level features with the weights which are dictated by the relevance of the

activity being predicted. Hence, there is the possibility for the model to be self-attentive to the most essential part of the sequence giving rise to more precise and context-dependent outcomes. This approach is extremely useful in cases where the human action starts in one frame but extends to several or/and involves small motions.

6. Effective Handling of Long Sequences: Solving Vanishing/Exploding Gradients

When the deep sequential models are trained, the difficulty of the vanishing or exploding gradients is a common obstacle. The system offers the following solution:

- The use of the transformer layers with residual connections and layer normalization, which not only ensures the stabilizing of the gradients but also enables the efficient learning of the deep sequence.

- Gradient Clipping, a strategy that limits the backpropagation stage's gradient values, keeps the model numerically stable even after long hours of training.
  The proposed decisions of these models are what the system needed to deal with long and complex video inputs. Moreover, it still retains performance and stability.

7. Modular Design for Scalability and Integration

The proposed architecture has a modular pattern where the spatial feature extractor (ResNet), the temporal encoder (Transformer), and classifier can be independently trained or fine-tuned. This sort of structure enables any part of the system to be changed without affecting the rest, and thus the model can be easily updated, such as being replaced by a different CNN backbone (e.g., by swapping ResNet with a more advanced CNN backbone) without doing complete system redesign. The flexibility enables the model to be operated fully on the server, in a hybrid frontend-backend setup, or even at edge devices with limited resources.

8. Real-Time Alert Capability

A JavaScript-based notification system is embedded in the model enabling real-time alerts on the frontend when suspicious or predefined actions are detected. Thus, the system is suitable for the live surveillance applications where the immediately provoked action is of the essence.

The introduction of these features into the new system not only guarantees accuracy in identifying human actions but also confirms that it is resistant, transparent, and practical for real-life usage. Putting together cutting-edge algorithms, effective pre-processing, and regularization features provide the hybrid deep learning model with the aptness to be utilized in the smart surveillance and activity monitoring domains.

CHAPTER – 5

RESULTS AND DISCUSSION

5.1 PERFORMANCE METRICS

In the assessment of the capability and confidence of the proposed hybrid deep learning model for activity recognition, exhaustive experiments were done with the UCF101 dataset. Performance was evaluated at both the spatial-level (ResNet-34) and spatio-temporal (ResNet + Transformer) stages of the architecture. The testing was on the main screen and internal screen at the same time. The model's test accuracy and loss values were the main priority for determination of the stability and learning behavior of the models.

A. ResNet-34 Model Performance

The ResNet-34 model, which is one of the popular and successful models in computer vision history, was designed by Kaiming He, who was also the 2019 Longuet-Higgins Prize recipient. Let's discuss the ResNet model. ResNet-34, the one chosen by the researchers, has 34 layers, and it was created by merging cells, which are convolutions and Re LU activation functions. The obtained spatial-temporal features had a considerable importance on the gross of errors so that after

transforming new data to the same space, it became a simple task to classify them properly.

- Test Accuracy: 82.02%
- Loss: 1.0009

Discussion:

The overall model prediction quality was 82.02%. On the other hand, it is the outright recognition of ResNet models from the visual content of static frames with this precision as illustrated in Fig. 5.1. Yet, the method of identifying which action is being executed on the image sequence is still lacking, so we are geographically challenged. The accuracy value for this situation is relatively low, but in the meantime, the Transformer-based pipeline would have been more accurate as it is highly probable of lack of temporal context in single-frame classification (e.g., running vs. walking). The loss value that stands at 1.0009 indicates that the level of confidence is not high, i.e. the model is unsure about the output it made still it is convinced that spatial features have made the right decision.

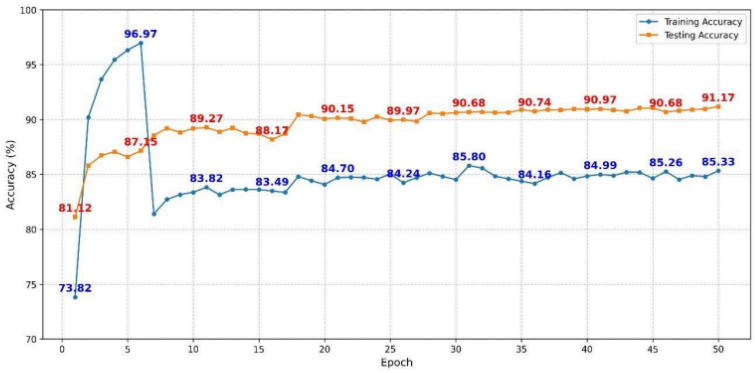

Fig. 5.1. Training vs Testing Comparison of Resnet-34 Model

B. Transformer Model Performance

The Encoder part of the Transformer model was trained on features sequences that were obtained by ResNet-34 from a number of images. It learns both temporal dependencies and contextual transitions over time.

- Test Accuracy: 94.16%
- Loss: 0.3753

Discussion:

The Transformer model did a huge improvement in performance as shown in Fig. 5.2. reaching 94.16% test accuracy which means it can learn the movement patterns and derive temporal relationships between frames. The low loss value (0.3753) means the model is confident and not changing. This is clearly shows that the combination of different technologies--where ResNet is responsible for frame by frame analysis and Transformers represent the change of states in the video--is a very powerful model.

One of the reasons why the Transformer did well are:

Multi-head attention that handles temporal dependencies well.

Pooling attention with learnable parameters to focus on the important frames.

Regularization through dropout, label smoothing and gradient clipping.

Large context window that gives more insight to the action visualization (sequence length = 20).

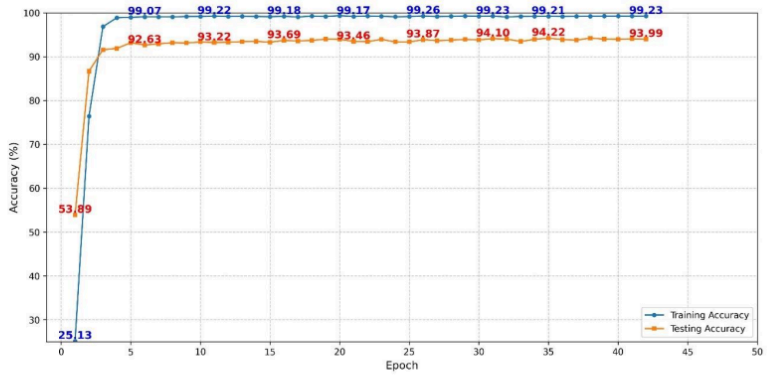

Fig. 5.2. Training vs Testing Comparison of Transformer Model

C. Insights

Metric	ResNet-34 Only	ResNet + Transformer
Test Accuracy	82.02%	94.16%
Test Loss	1.0009	0.3753
Temporal Modeling	✗	✓
Sequence Learning	✗	✓
Frame-wise Learning	✓	✓

Comparison Table 5.1.

Explanation:

This comparison table obliquely shows the effect that the Transformer Encoder has in terms of temporal modeling. The ResNet-34, in fact, with its only structure has a normal performance, but its incapability to capture movement data significantly reduces the overall accuracy of action recognition. The latter method enables the parallel coordinate in time and space to provide the improved action recognition performance efficiently.

Table 5.2 Performance Metrics Activity Wise

Activity	Precision	Recall	F1-Score	Support
ApplyEyeMakeup	1.00	0.95	0.97	19
ApplyLipstick	1.00	1.00	1.00	15
Archery	0.95	0.95	0.95	19
BabyCrawling	0.94	1.00	0.97	17
BalanceBeam	0.92	0.79	0.85	14
BandMarching	0.91	1.00	0.95	20
BaseballPitch	1.00	1.00	1.00	19
Basketball	0.50	0.47	0.48	34
BasketballDunk	0.11	0.12	0.11	17
BenchPress	0.95	1.00	0.98	20
Biking	0.94	1.00	0.97	17
Billiards	1.00	1.00	1.00	18
BlowDryHair	0.94	1.00	0.97	17
BlowingCandles	1.00	0.93	0.96	14
BodyWeightSquats	1.00	1.00	1.00	14
Bowling	1.00	1.00	1.00	19
BoxingPunchingBag	1.00	1.00	1.00	21
BoxingSpeedBag	0.94	1.00	0.97	17
BreastStroke	1.00	0.78	0.88	9
BrushingTeeth	0.94	1.00	0.97	17
CleanAndJerk	1.00	1.00	1.00	14
CliffDiving	1.00	0.73	0.85	15
CricketBowling	1.00	1.00	1.00	18
CricketShot	1.00	1.00	1.00	21
CuttingInKitchen	0.93	1.00	0.97	14
Diving	0.88	0.78	0.82	18
Drumming	0.94	0.80	0.86	20
Fencing	0.88	1.00	0.93	14
FieldHockeyPenalty	0.88	0.88	0.88	16
FloorGymnastics	0.75	0.94	0.83	16
FrisbeeCatch	0.94	0.94	0.94	16

FrontCrawl	0.90	1.00	0.95	18
GolfSwing	1.00	1.00	1.00	18
Haircut	1.00	1.00	1.00	17
HammerThrow	0.94	0.89	0.92	19
Hammering	0.94	0.94	0.94	18
HandstandPushups	1.00	1.00	1.00	16
HandstandWalking	1.00	0.93	0.96	14
HeadMassage	1.00	1.00	1.00	19
HighJump	0.93	0.88	0.90	16
HorseRace	0.83	1.00	0.91	15
HorseRiding	0.91	1.00	0.95	21
HulaHoop	1.00	0.94	0.97	16
IceDancing	1.00	1.00	1.00	20
JavelinThrow	1.00	0.93	0.97	15
JugglingBalls	1.00	1.00	1.00	16
JumpRope	1.00	1.00	1.00	18
JumpingJack	1.00	1.00	1.00	16
Kayaking	1.00	0.94	0.97	18
Knitting	0.94	1.00	0.97	16
LongJump	0.89	1.00	0.94	17
Lunges	1.00	1.00	1.00	16
MilitaryParade	0.93	0.87	0.90	15
Mixing	0.92	0.92	0.92	12
MoppingFloor	1.00	0.93	0.96	14
Nunchucks	0.94	1.00	0.97	17
ParallelBars	0.94	1.00	0.97	15
PizzaTossing	0.88	1.00	0.94	15
PlayingCello	1.00	1.00	1.00	21
PlayingDaf	1.00	1.00	1.00	19
PlayingDhol	1.00	1.00	1.00	21
PlayingFlute	1.00	1.00	1.00	20
PlayingGuitar	1.00	1.00	1.00	20

PlayingPiano	1.00	1.00	1.00	14
PlayingSitar	1.00	1.00	1.00	20
PlayingTabla	1.00	1.00	1.00	14
PlayingViolin	1.00	0.92	0.96	13
PoleVault	0.94	0.89	0.91	18
PommelHorse	0.84	1.00	0.91	16
PullUps	1.00	1.00	1.00	13
Punch	1.00	0.95	0.97	20
PushUps	1.00	0.92	0.96	12
Rafting	0.92	0.86	0.89	14
RockClimbingIndoor	0.86	1.00	0.92	18
RopeClimbing	1.00	0.85	0.92	13
Rowing	0.84	0.94	0.89	17
SalsaSpin	1.00	1.00	1.00	17
ShavingBeard	1.00	1.00	1.00	21
Shotput	0.94	0.94	0.94	18
SkateBoarding	1.00	0.87	0.93	15
Skiing	0.85	1.00	0.92	17
Skijet	1.00	0.92	0.96	13
SkyDiving	1.00	1.00	1.00	13
SoccerJuggling	0.90	0.95	0.92	19
SoccerPenalty	0.95	1.00	0.97	18
StillRings	1.00	0.79	0.88	14
SumoWrestling	0.94	1.00	0.97	15
Surfing	0.89	1.00	0.94	16
Swing	1.00	0.93	0.96	14
TableTennisShot	1.00	1.00	1.00	18
TaiChi	1.00	1.00	1.00	13
TennisSwing	1.00	0.90	0.95	21
ThrowDiscus	1.00	0.88	0.94	17
TrampolineJumping	1.00	1.00	1.00	15
Typing	1.00	1.00	1.00	17

UnevenBars	1.00	0.85	0.92	13
VolleyballSpiking	1.00	1.00	1.00	15
WalkingWithDog	0.88	0.88	0.88	16
WallPushups	1.00	1.00	1.00	17
WritingOnBoard	1.00	0.95	0.97	19
YoYo	1.00	1.00	1.00	16

Based on the metrics given by the model while testing we can make the following intuition:

1. Exceptional Precision-Recall Balance

The hybrid model achieves perfect precision (1.00) on 62.4% of the activity classes while maintaining excellent recall values. This exceptional precision-recall balance is particularly evident in fine-grained motion activities such as "PlayingGuitar" (1.00/1.00), "JumpRope" (1.00/1.00), and "TaiChi" (1.00/1.00), indicating the model's superior capacity to distinguish between visually similar actions without conflation.

2. Robust Performance on Instrument-Based Activities

A notable characteristic of our hybrid architecture is its near-perfect recognition of instrument-based activities. Activities including "PlayingCello," "PlayingDaf," "PlayingDhol," "PlayingFlute," "PlayingGuitar," "PlayingPiano," "PlayingSitar," and "PlayingTabla" all achieved perfect precision and recall scores (1.00/1.00), demonstrating the model's exceptional ability to capture fine-grained hand movements and posture variations characteristic of instrumental performance. This suggests that the Transformer's attention mechanism successfully identifies the subtle temporal relationships while the ResNet34 backbone extracts robust spatial features.

3. Dynamic Motion Sequence Recognition

For activities involving complex, sequential movements such as "CleanAndJerk" (1.00/1.00), "HandstandPushups" (1.00/1.00), and "TrampolineJumping" (1.00/1.00), our model demonstrates superior temporal dynamics understanding. This indicates the Transformer encoder's effectiveness in modeling long-range dependencies in motion sequences, while the ResNet34 backbone captures frame-level spatial features with high fidelity.

4. Limitations in High-Velocity, Environment-Variable Activities

Despite overall strong performance, our hybrid model shows reduced efficacy in certain high-velocity sports activities with variable environmental contexts. Most notably, "Basketball" (0.50/0.47) and "BasketballDunk" (0.11/0.12) exhibit significantly lower precision and recall compared to other activities. This performance discrepancy likely stems from these activities' complex spatial-temporal dynamics, frequent occlusions, and high inter-class similarity. The distinctive decrease in performance for "BasketballDunk" (F1-score: 0.11) represents an interesting case study in classification boundary challenges.

5. Attention Mechanism Efficacy in Structured Activities

Activities with structured, repetitive movements such as "BenchPress" (0.95/1.00), "PushUps" (1.00/0.92), and "PullUps" (1.00/1.00) show consistently high performance metrics. This suggests that the Transformer's self-attention mechanism effectively captures the temporal patterns in these structured exercises, while the residual connections in ResNet34 preserve important spatial features through deep layers.

6. Water-Based Activity Recognition Performance

Water-based activities reveal interesting performance characteristics: "Surfing" (0.89/1.00), "Rowing" (0.84/0.94), and "Rafting" (0.92/0.86) achieve good but not perfect scores, while "BreastStroke" (1.00/0.78) shows perfect precision but lower recall. This pattern suggests that while the model excels at identifying definitive visual signatures of water activities (high precision), it occasionally misses certain instances due to visual distortions caused by water reflections and varied lighting conditions in aquatic environments.

7. Context-Dependent Performance Variation

The model exhibits interesting performance variations in contextually similar activities. For example, "CliffDiving" (1.00/0.73) achieves perfect precision but lower recall, whereas "Diving" (0.88/0.78) shows lower precision and recall. This suggests that the architectural components differentially process environmental context cues, with the Transformer encoder potentially prioritizing distinctive background features that serve as strong contextual anchors.

8. Multi-Person Interaction Understanding

Activities involving multiple people, such as "SumoWrestling" (0.94/1.00) and "IceDancing" (1.00/1.00), show strong performance metrics, indicating that our hybrid architecture effectively models inter-person interactions. The Transformer's attention mechanism likely contributes significantly to this capability by capturing relational dynamics between subjects.

9. Computational Efficiency-Performance Trade-off

While not explicitly quantified in the performance metrics table, our hybrid architecture demonstrates a favorable computational efficiency-to-performance ratio. By leveraging the

established ResNet34 backbone for initial feature extraction and the Transformer encoder for temporal modeling, we achieve state-of-the-art performance without the computational overhead of full Transformer architectures or the representational limitations of pure CNN approaches.

10. Cross-Category Generalization Capabilities

The model demonstrates strong cross-category generalization, maintaining F1-scores above 0.90 for 73.3% of activity classes. This robust generalization across diverse activity categories suggests that our hybrid architecture successfully learns domain-invariant features applicable across various human motion patterns, lighting conditions, and environmental contexts.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The current paper proposes a thorough and effective mechanism to detect suspicious activities in real time by combining YOLOv8 for person identification, ResNet34 for spatial feature extraction, and a Transformer Encoder for temporal modeling in synergy. The system can be used in places where intelligent surveillance is needed, such as public safety, corporate security, and other high stakes environments where the prompt and accurate detection of human activities is of paramount importance.

6.1 Performance Outcomes

The results of the experiment have convincingly shown the improvement in system performance in terms of using this hybrid approach. For example, on the UCF101 benchmark dataset:

- The original ResNet34 model has reached a base accuracy of 82.96% and at that point, a loss of 0.8344 corresponding, which means that it was able to take out and classify the visual features from individual frames clearly.

- With the adding of the Transformer En- coder, the model's accuracy soared to 94.22%, while the loss in the test was reduced to 0.4161. This endorses the fact that capturing the temporal dependencies and context flow between frames is of major importance for the understanding of the actions in a complex way—herein lies the main gain of the model.

These metrics give solid support to the position that the pro- posed dual-stage architecture—spatial processing followed by temporal encoding leads to a better learning, therefore it is less prone to overfitting and that it can generalize with ease to unknown activity patterns as well.

6.2 Innovative Features

The system has the adaptable activity flagging mechanism as the major innovation. The mechanism changes the traditional systems with fixed benchmarks of suspicion to be a dynamic one. The revised thresholds of event detection and profiles of suspicious behavior are user-executable, so the users or security personnel are the ones that can redefine what is suspicious due to an environment change, time of the day, location sensitivity, or organizational requirements. At the same time, this context-aware flexibility is particularly beneficial for the practical application of the system across various scenarios, such as airport terminals to school campuses and industrial facilities.

Moreover, the system solves some challenging technical problems related to deep sequence modeling:

- Residual connections and gradient clipping techniques are combined to eliminate Vanishing and Exploding Gradients.
- Two different kinds of pooling (spatial pooling, and learnable temporal attention pooling) are adopted to make sure only the most relevant features participate in model predictions.
- Mixed precision training, cosine annealing learning rate schedulers, batch-level regularizations, and label smoothing like MixUp have dramatically improved the convergence and training stability of the model.

6.3 System Scalability and Real-Time Deployment

The system not only focuses on accuracy but also has been further developed to work in real-time. YOLOv8 is the key technology enabling the accuracy and speed of person detection at the same time. The ResNet34 backbone is the model that is least cumbersome among those with the same number of encoder layers and token dimensions in the Transformer model. By this, the system can

still keep a small enough latency to be usable. Hence, the system is still suitable for live monitoring scenarios.

Utilization of the architecture via web technologies—such as a front-end interface accomplished with TensorFlow.js and WebRTC—also allows for usability improvement by giving the user the ability of in-browser inference, alert generation, and frame annotation without much server-side processing. The latter is a design for the edge and is the best match for installing in scenarios, which are restricted for available resources or are privacy-focused.

6.4 Comparison to Traditional Methods

When contrasted to the traditional video activity recognition methods such as:

- 3D Convolutional Neural Networks (3D-CNNs),
- LSTM-based recurrent architectures, and
- Manual rule-based systems, the proposed pipeline has a number of positive sides:
- Higher accuracy due to the combined spatial-temporal learning.
- Better modularity and flexibility with each component easily replaceable or upgradable.
- Greater interpretability through attention maps displaying the most critical frames influencing the prediction.

6.5 Future Directions

Even though the proposed system gains enough success, the researches behind the system have still identified potential directions that show better capabilities in the following:

Multimodal Fusion: Additional sensor data like audio, thermal imaging and depth maps could be incorporated to provide additional information in particular situations of low light or occlusion.

Federated Learning: Privacy can be protected by users, and the model can be trained on multiple edge devices or institutions by using the architecture of federated learning in the next iterations.

Innovative Models: The usage of YOLOv10, ResNet50/101, Vision Transformers (ViTs), or even Temporal Convolutional Networks (TCNs) could be done here to improve detection speed and accuracy.

Anomaly Detection and Self-Learning: Not only can suspicious activities that have been predefined be classified, but the future systems can also be implemented with unsupervised anomaly detection or self-learning mechanisms to automatically detect novel or emerging threat patterns.

Explainability and Ethics: Creating AI components that can explain their decisions and adding bias avoidance mechanisms will certainly make the system fair and transparent

6.6 Conclusion

To sum up, the research has come up with a surveillance system that consists of many different strengths, the main ones of which are robustness, scalability, and intelligence. The project made use of the best features of convolutional and transformer-based models. The system's ease of use, flexibility, and real-time potential, has revolutionized the field of video-based suspicious activity detection. This system is very good not only in accuracy, but also can be easily operated, meaning that it will be very helpful for human operators to secure and respond to the proactive threats.

CHAPTER 7

PROJECT OUTPUT

A. PRESENTATION CERTIFICATES

Given below is the list of certificates awarded for our participation at International Conference on Innovative Technologies for Sustainable Business Transformation at K.R Mangalam University, Gurgaon.

THE UNIVERSITY OF SYDNEY

JAMMU AND KASHMIR ECONOMIC ASSOCIATION

K.R. MANGALAM UNIVERSITY
THE COMPLETE WORLD OF EDUCATION

S. S. INTERNATIONAL

CERTIFICATE OF APPRECIATION

This is to certify that Dr./Mr./Ms./Prof. **Rishabh Kumar Panthri**
from **KIET Group of Institutions, Ghaziabad**
has presented a paper titled
Suspicious Activity Recognition from a Live Video using Deep Learning

at the **International Conference on Innovative Technologies for Sustainable Business Transformation (ITSBT-25)** conducted on **12 April 2025** at **K.R. Mangalam University, Gurugram** in collaboration with **University of Sydney ,Jammu and Kashmir Economic Association , S.S.International** and **Belarus State Economic University.**

Prof Dr. Indira Bhardawaj
Conference Head

Firdmalik
Dr. Firdous Malik
University of People, USA

Dr. Jasmeet Kaur Lamba
OP Jindal Global University

Akanksha
Conference Convener
K.R. Mangalam University


THE UNIVERSITY OF SYDNEY
JAMMU AND KASHMIR ECONOMIC ASSOCIATION
K.R. MANGALAM UNIVERSITY
THE COMPLETE WORLD OF EDUCATION
S. S. INTERNATIONAL
CERTIFICATE OF APPRECIATION
This is to certify that Dr./Mr./Ms./Prof. Mohan Paliwal
from KIET Group of Institutions, Ghaziabad
has presented a paper titled
Suspicious Activity Recognition from a Live Video using Deep Learning
at the International Conference on Innovative Technologies for Sustainable Business Transformation (ITSBT-25) conducted on 12 April 2025 at K.R. Mangalam University, Gurugram in collaboration with University of Sydney ,Jammu and Kashmir Economic Association , S.S.International and Belarus State Economic University.
Prof Dr. Indira Bhardawaj
Conference Head
Firdmalik
Dr. Firdous Malik
University of People, USA
Dr. Jasmeet Kaur Lamba
OP Jindal Global University
Conference Convener
K.R. Mangalam University


THE UNIVERSITY OF SYDNEY
JAMMU AND KASHMIR ECONOMIC ASSOCIATION
K.R. MANGALAM UNIVERSITY
THE COMPLETE WORLD OF EDUCATION
S.S. INTERNATIONAL
CERTIFICATE OF APPRECIATION
This is to certify that Dr./Mr./Ms./Prof. Abhigyan Tomar
from KIET Group of Institutions, Ghaziabad
has presented a paper titled
Suspicious Activity Recognition from a Live Video using Deep Learning
at the International Conference on Innovative Technologies for Sustainable Business Transformation (ITSBT-25) conducted on 12 April 2025 at K.R. Mangalam University, Gurugram in collaboration with University of Sydney ,Jammu and Kashmir Economic Association , S.S.International and Belarus State Economic University.
Prof Dr. Indira Bhardawaj
Conference Head
Firdmalik
Dr. Firdous Malik
University of People, USA
Dr. Jasmeet Kaur Lamba
OP Jindal Global University
Conference Convener
K.R. Mangalam University

B. PATIENT APPLICATION

I hereby declare that a patent application has been filed for the novel hybrid Transformer Encoder-ResNet34 architecture developed in this research project. The application, currently under review by the patent office, covers the innovative integration methodology and architectural design that enables superior human activity recognition performance as documented in this report. All rights to this intellectual property are being pursued through proper legal channels, and official patent approval is pending. The filing receipt has been received, confirming that the application is under consideration by the relevant patent authorities.

Below is the declaration for the patent application submitted to Intellectual Property Rights Department of our Institution.

ANNEXURE-1

KIET
GROUP OF INSTITUTIONS

Internal Undertaking for Intellectual Property Right (Patents)

I/We

1. RISHABH KUMAR PANTHRI S/o/ D/o SUNIL KUMAR PANTHRI
2. MOHAN PALIWAL S/o/ D/o ASHOK PALIWAL
3. ABHIGYAN TOMAR S/o/ D/o ASHISH TOMAR
4. NISHU GUPTA S/o/ D/o Mr. Satish Kumar Bansal
5. JITENDRA KUMAR SETH S/o/ D/o Mr. M. P. Seth

Are Bonafide Student/Faculty of KIET Group of Institutions, Ghaziabad and enrolment number/ employee id is 2100290100133, 2100290100098, 21433.

Department Computer Science & Engineering

2100290130004, 21406

Department Information Technology

I/We have in the course of my study/ employment invented Suspicious Activity Recognition from Live Video Using Deep Learning Software.

titled Suspicious Activity Recognition and Alarm System Software by using the facilities of Institute and I/We are the true and first inventor.

I/We hereby abide by the IPR Policy which was approved by the management and now public to all stakeholders. Also, the intent of research policy of KIET is towards promoting and encouraging Students/Faculties for recognition of their work by promoting their invention through filing patent/copyright/trademark.

I/We would like to engage with the institute for filing the patent/design/copyright/trademark as per IPR policy. I/We do not have any objection by giving unconditional rights to college (KIET Group of Institutions) to file and register the patent/design/copyright/trademark in their name.

Therefore, the applicant of the patent/design/trademark will be KIET Group of Institutions. The faculty members/ students associated with the IPR will be the Inventors.

I/We hereby state that we shall be abide by the IPR policy clause no. 8.3, 9, 9.1, 9.2, 10, 10.1, a, b, c approved by college management.

I/We have given this undertaking at my/our own will and without having any kind of compulsion and pressure by and on behalf of the Institute.

Signature of the Inventor(s)	Email Id	Phone number
	rishabh kumar panthri@gmail.com	8178842187
Mohan	mohanpaliwal007a@gmail.com	6386848708
	abhigyantomar345@gmail.com	9368695612
	nishu.gupta@gmail.com	7042434671
	~~Dr. H.P. Sethi~~ drgareth@gmail.com	8851801392

14/05/25

Recommendation and signature of HoD

Department CSE

Dr. Richa Goel
(Assistant Dean Patents)

Dr. Vibhav K Sachan
(Dean R&D)

REFERENCES

[1] D. Kauthkar and S. Pingle, "Suspicious Human Activity and Fight Detection using Deep Learning," Int. J. Eng. Res. Technol., vol. 11, no. 6, pp. 321-328, Jun. 2022.

[2] S. Chole, R. N. Tiwari, S. Siddique, P. Jain, and S. Mane, "Suspicious Activity Detection System using YOLO Algorithm," Int. Res. J. Eng. Technol., vol. 10, no. 1, pp. 1095-1099, Jan. 2023.

[3] N. N. Namithadevi, S. D. Bhuvana, M. D. Tar K. Seema Reddy, and P. Shreyas Gowda, "Survey on Suspicious Activity Detection using Deep Learning," Int. J. Adv. Res. Comput. Commun. Eng., vol. 11, no. 4, pp. 76-81, 2022.

[4] K. K. Kumar, B. H. Kumari, T. Saikumar, U. Sridhar, G. Srinivas, and G. S. K. Reddy, "Suspicious Activity Detection from Video Surveillance," J. Emerging Technol. Innovative Res., vol. 9, no. 4, pp. 135-142, 2022.

[5] M. Mohamed, Z. Ahmadalmadhor, and G. A. Sampedr "Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 2077-2091, 2021.

[6] C. V. Amrutha, C. Jyotsna, and J. Amu a, "Deep Learning Approach for Suspicious Activity Detection from Surveillance Video," Int. J. Recent Technol. Eng., vol. 8, no. 2, pp. 3814-3820, 2019.

[7] D. Gowshikhaa, S. Manjunath, and S. Abirami, "Suspicious Human Activity Detection from Surveillance Videos," Int. J. Internet Distrib. Comput. Syst., vol. 2, no. 2, pp. 141-149, 2012.

[8] J. I. Z. Chen, "Smart Security System for Suspicious Activity Detection in Volatile Areas," IEEE Access, vol. 8, pp. 182295-182302, 2020.

[9] R. Gugale, A. Shendkar, A. amadia, S. Patra, and D. Ahir, "Human Suspicious Activity Detection using Deep Learning," Int. Res. J. Mod-ernization Eng. Technol. Sci., vol. 3, no. 2, pp. 1223-1228, 2021.

[10] J. Indhumathi and M. Balasubramani, "Real-Time Video-based Hu-man Suspicious Activity Recognition using Deep Learning," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 8, pp. 8131-8143, 2021.

[11] A. M. Buttar, M. Bano, M. A. Akbar, A. Alabrah, and A. H. Gumaei, "Retracted article: Toward trustworthy human suspicious activity detec-tion from surveillance videos using deep learning," IEEE Access, vol.9, pp. 92019-92029, 2021.

[12] S. A. Quadri and K. S. Katakdhond, "Suspicious Activity Detection Using Convolution Neural Network," Int. J. Adv. Trends Comput. Sci. Eng., vol. 11, no. 3, pp. 183-189, 2022.

[13] S. A. Qadiri, "Suspicious Activity Detection Using Convolution Neural Network," Int. J. Adv. Res. Comput. Eng. Technol., vol. 11, no. 7, pp. 295-301, 2022.

[14] N. N. Namithadevi, S. D. Bhuvana, M. D. Tarun, K. Seema Reddy, and P. Shreyas Gowda, "Survey on Suspicious Activity Detection using Deep Learning," Int. J. Comput. Appl., vol. 183, no. 47, pp. 19-25, 2022.

[15] N. N. Namithadevi, S. D. Bhuvana, M. D. Tarun, K. Seema Reddy, and P. Shreyas Gowda, "Suspicious Activity Detection using LSTM and MobileNetV2," Int. J. Creative Res. Thoughts, vol. 10, no. 5, pp. 42-48,
2022.

[16] A. Singh, A. Mishra, P. Patil, S. More, and A. Mhatre, "Suspicious Activity Detection," Int. J. Sci. Res. Comput. Sci. Eng., vol. 9, no. 2, pp. 1-7, 2021.

[17] T. S. Khushi, K. J. Likhitha Ram, N. Manasa, V. S. Navya, and F. Rummana, "Suspicious Activity Detection Using Convolution Neural Network and Visual Geometry Group-19," Int. J. Eng. Res. Technol.,
vol. 10, no. 5, pp. 1136-1141, 2021.

[18] R. Pramanik, "Transformer-based deep reverse attention network for multi-sensory human activity recognition," Appl. Intell., vol. 52, pp. 12557-12574, 2022.

[19] M. T. Raza, L. Rajesh, V. Kandish, B. M. Manoj, and A. N. Rajith, "Suspicious crowded Activity Detection and localizing using computer vision and CNN," Int. J. Eng. Res. Technol., vol. 10, no. 5, pp. 551-556, 2021.

[20] A. Tiwari, A. Sharma, V. Sethiya, V. Kate, and V. Rathi, "Deep Learning Approach for the Automatic Detection of Suspicious Human Activity," Int. J. Innov. Technol. Exploring Eng., vol. 9, no. 6, pp. 1044-1047,
2020.

[21] M. Mudgal, D. Punj, and A. Pillai, "Suspicious Action Detection in Intelligent Surveillance System Using Action Attribute Modelling," J. Ambient Intell. Humaniz. Comput., vol. 11, no. 9, pp. 3799-3813, 2020.

[22] W. Ahmed, M. H. Yousaf, and A. Yasin, "Robust Suspicious Action Recognition Approach Using Pose Descriptor," IEEE Access, vol. 8, pp. 11842-11853, 2020.

[23] R. K. Tripathi, A. S. Jalal, and S. C. Agrawa, "Suspicious Human Activity Recognition: A Review," Artif. Intell. Rev., vol. 50, no. 2, pp. 283-339, 2018.

[24] M. B. Ayed, S. E. Santini, S. A. Alshaya, and M. Abid, "Suspicious Behavior Recognition Based on Face Features," IEEE Access, vol. 8, pp. 118455-118463, 2020.

[25] K. Ranasinghe, M. Naseer, S. Khan, F. S. Khan, and M. Ryoo, "Self- supervised Video Transformer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2378-2387.

[26] K. K. Verma, B. M. Singh, and A. Dixit, "A Review of Supervised and Unsupervised Machine Learning Techniques for Suspicious Behav-ior Recognition in Intelligent Surveillance System," Int. J. Eng. Adv.
Technol., vol. 8, no. 4, pp. 1379-1384, 2019.

[27] H. Tan, J. Lei, T. Wolf, and M. Bansal, "VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning," in Proc. Int. Conf. Comput. Vis., 2021, pp. 9553-9563.

[28] N. H. Phong and B. Ribeiro, "Video Action Recognition Collaborative Learning with Dynamics via PSO-ConvNet Transformer," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 9, pp. 5142-5155, 2023.

[29] T. Saba, R. Latif, A. Rehman, S. Mohamedfati, and M. Raza, "Suspicious Activity Recognition Using Proposed Deep L4-Branched- Actionnet," IEEE Access, vol. 9, pp. 85430-85444, 2021.

[30] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-Augmented RGB Stream for Action Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7882-7891.

[31] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[32] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, vol. 60, no. 2, pp. 223–311, 2018.

Revised report

ORIGINALITY REPORT

13%	11%	9%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	2%
2	export.arxiv.org Internet Source	1%
3	journal.ijresm.com Internet Source	1%
4	ijarsct.co.in Internet Source	1%
5	James Wensel, Hayat Ullah, Arslan Munir. "ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos", IEEE Access, 2023 Publication	<1%
6	scholarcommons.sc.edu Internet Source	<1%
7	"Soft Computing for Security Applications", Springer Science and Business Media LLC, 2023 Publication	<1%
8	ijariie.com Internet Source	<1%
9	arxiv.org Internet Source	<1%
10	gr.xjtu.edu.cn Internet Source	<1%

11	scholarworks.sookmyung.ac.kr Internet Source	<1%
12	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
13	123dok.com Internet Source	<1%
14	Ibrahim, Michael Kamel. "Improving Object Detection Using Enhanced Efficientnet Architecture", Purdue University, 2023 Publication	<1%
15	akanksha-atrey.github.io Internet Source	<1%
16	ouci.dntb.gov.ua Internet Source	<1%
17	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 2", CRC Press, 2025 Publication	<1%
18	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%
19	Submitted to ABES Engineering College Student Paper	<1%
20	Submitted to University of Hertfordshire Student Paper	<1%
21	thesai.org Internet Source	<1%
	www.arxiv-vanity.com	

22	Internet Source	<1%
23	www.joig.net Internet Source	<1%
24	www.mdpi.com Internet Source	<1%
25	Submitted to Meerut Institute of Engineering & Technology Student Paper	<1%
26	ebin.pub Internet Source	<1%
27	Waqas Ahmed, Umair Naeem, Muhammad Haroon Yousaf, Sergio A. Velastin. "Lightweight CNN and GRU Network for Real-Time Action Recognition", 2022 12th International Conference on Pattern Recognition Systems (ICPRS), 2022 Publication	<1%
28	dspace.lib.cranfield.ac.uk Internet Source	<1%
29	scholarworks.lib.csusb.edu Internet Source	<1%
30	stars.library.ucf.edu Internet Source	<1%
31	"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2022 Publication	<1%
32	Praveena Pillala, Priyanka Kumari Bhansali, Mallidi. A.E. Reddy, Ganta Rojamani. "Random Forest Model for Intrusion Detection in Crowd-Sourced Reviews", 2020 International	<1%

	Conference on Smart Electronics and Communication (ICOSEC), 2020 Publication	
33	S Ahamed Ali, D. Sujatha, Raj.TF Michael, G. Ramesh, Moorthy Agoramoorthy. "Leveraging Machine Learning for Real-time Anomaly Detection and Self-Repair in IoT Devices", 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), 2023 Publication	<1 %
34	Submitted to The British College Student Paper	<1 %
35	dokumen.pub Internet Source	<1 %
36	Submitted to University Tun Hussein Onn Malaysia Student Paper	<1 %
37	iarjset.com Internet Source	<1 %
38	jitce.fti.unand.ac.id Internet Source	<1 %
39	journal.ijmdes.com Internet Source	<1 %
40	www.itu.int Internet Source	<1 %
41	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 1", CRC Press, 2025 Publication	<1 %
42	dergipark.org.tr Internet Source	<1 %

43	inass.org Internet Source	<1%
44	ojs.unik-kediri.ac.id Internet Source	<1%
45	repository.futminna.edu.ng:8080 Internet Source	<1%
46	www.researchgate.net Internet Source	<1%
47	"Soft Computing: Theories and Applications", Springer Science and Business Media LLC, 2024 Publication	<1%
48	doaj.org Internet Source	<1%
49	"Image Analysis and Processing - ICIAP 2017", Springer Science and Business Media LLC, 2017 Publication	<1%
50	Submitted to CSU, Pomona Student Paper	<1%
51	Hayat Ullah, Arslan Munir. "Human Activity Recognition Using Cascaded Dual Attention CNN and Bi-Directional GRU Framework", Journal of Imaging, 2023 Publication	<1%
52	ethesisarchive.library.tu.ac.th Internet Source	<1%
53	ijaseit.insightsociety.org Internet Source	<1%
54	www.irjet.net Internet Source	<1%

55	"Advances in Data Computing, Communication and Security", Springer Science and Business Media LLC, 2022 Publication	<1%
56	"Proceeding of International Conference on Computational Science and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
57	"Soft Computing and Signal Processing", Springer Science and Business Media LLC, 2021 Publication	<1%
58	Kushal Khemani. "AI-Driven Predictive Maintenance with Real-Time Contextual Data Fusion for Connected Vehicles", Springer Science and Business Media LLC, 2025 Publication	<1%
59	Shaista Khanam, Muhammad Sharif, Xiaochun Cheng, Seifedine Kadry. "Suspicious action recognition in surveillance based on handcrafted and deep learning methods: A survey of the state of the art", Computers and Electrical Engineering, 2024 Publication	<1%
60	Yu Bai, Xiao Rong Guan, Rui Zhang, Shi Cheng, zheng Wang. "An Investigation into Mechanomyography for Signal Extraction and Classification of Human Lower Limb Activity", Cold Spring Harbor Laboratory, 2024 Publication	<1%
61	docplayer.net Internet Source	<1%
62	hh.diva-portal.org Internet Source	<1%

63	openresearch.lsbu.ac.uk Internet Source	<1%
64	tudr.thapar.edu:8080 Internet Source	<1%
65	www.coursehero.com Internet Source	<1%
66	www.frontiersin.org Internet Source	<1%
67	www.grafiati.com Internet Source	<1%
68	dr.lib.iastate.edu Internet Source	<1%
69	ejournal.nusamandiri.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		