Optimized Multimodal Human Activity and Anomaly Detection

Arya Dinachandran, Prof Rameez Mohammed A.

Department of Computer Science and Engineering, College of Engineering, Trivandrum, India

DOI: https://doi.org/10.51244/IJRSI.2026.1303000197

Received: 24 March 2026; Accepted: 30 March 2026; Published: 15 April 2026

ABSTRACT

Human activity recognition and anomaly detection play a crucial role in applications such as intelligent surveillance, healthcare monitoring, and smart environments. This study proposes a optimized multimodal framework that integrates video data and IMU sensor signals to differentiate normal and abnormal human activities efficiently. Each data stream is processed independently using optimized, low-complexity deep learning models, and predictions are combined at the decision level to enhance accuracy while avoiding the overhead of feature fusion. Model Optimization reduces memory usage, model size, and inference time, enabling deployment on edge devices such as smart cameras, smartwatches, and smartphones. The system is evaluated on various activities, including fall, driver drowsiness as well as normal activity like walk, run etc, under diverse lighting conditions, sensor placements, and environmental variations to ensure robust performance. The framework emphasizes real-time monitoring and low-latency response, providing a scalable and practical solution for continuous anomaly detection. Future work may extend the system to additional modalities and incremental learning for improved adaptability.

Index Terms: Anomaly Detection, Multimodal Learning, Optimization, Deep Learning, IMU Signals, Video Analysis, Edge Computing.

INTRODUCTION

Human Activity Recognition (HAR) and anomaly detection are increasingly critical in modern intelligent systems, supporting applications such as public surveillance, elderly care, rehabilitation, and workplace safety. HAR focuses on identifying and categorizing human actions, while anomaly detection aims to recognize unusual or unsafe behaviors that may indicate potential hazards. As environments become more automated and context-aware, there is a growing demand for systems capable of accurately detecting human activities and anomalies in real time, reducing the need for continuous human supervision and enhancing safety, efficiency, and situational awareness.

Previous approaches often relied on a single modality, such as either video or sensor data, which limited reliability under varying conditions, including dynamic lighting, occlusions, or irregular sensor placement. Multimodal learning methods, which combine data from multiple sources, have emerged to address these limitations. Integrating visual information from video with motion data from Inertial Measurement Unit (IMU) sensors enables the system to leverage both appearance-based and movement-based cues, improving robustness and adaptability for diverse deployment scenarios.

This work presents a optimized multimodal framework for human activity recognition and anomaly detection, where video and IMU data are processed independently using lightweight, optimized deep learning models. Temporal pat- terns from video are extracted using LSTM Autoencoders, while motion features from IMU signals are modeled using 1D CNN–GRU networks. Predictions are generated at the decision level rather than through feature fusion, allowing each model to operate efficiently on edge devices such as smart cameras, wearables, and smartphones. Optimization further reduces model size, memory requirements, and inference time, enabling low-latency, real-time performance without compromising accuracy.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

The main contributions of this research include the design of an independent, multimodal framework optimized for edge deployment, the application of Optimization techniques for efficient real-time monitoring, and the evaluation of the system across multiple anomaly scenarios such as falls, driver drowsiness under diverse environmental conditions. The objectives are to improve reliability, scalability, and efficiency while minimizing computational and storage requirements.

The remainder of the paper is organized as follows: Section 2 presents the related work and highlights the advantages and limitations of previous methods. Section 3 details the proposed methodology, including the network architectures, Optimization strategy, and multimodal processing. Section 4 presents the experimental setup, results, system requirements, and performance analysis. Section 5 discusses the conclusions, limitations, and future research directions.

TRADITIONAL METHODS AND THEIR CHALLENGES

Human Activity Recognition (HAR) and anomaly detection have been widely studied, with conventional approaches offering foundational insights but facing significant limitations in scalability, multimodal integration, and deployment on edge devices.

Traditional Machine Learning Approaches: Early HAR systems primarily relied on traditional machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests [1].These methods depended on handcrafted features extracted from accelerometer and gyroscope data, including statistical measures and domain-specific descriptors. While effective on small or controlled datasets, these approaches exhibited limited generalization to complex or dynamic activities. They were also sensitive to noisy data, sensor drift, and variations in human motion, and lacked scalability for integrating multiple modalities, such as combining IMU and video data, which restricted their applicability in real-world scenarios.

Single-Modality Systems: Many early HAR and anomaly detection frameworks operated using either video or sensor data exclusively [2]. Vision-based methods employed techniques such as optical flow, background subtraction, and silhouette analysis to classify activities. Although capable of capturing appearance-based cues, their performance declined under challenging conditions, including poor lighting, occlusions, and unstable camera perspectives. Sensor-based methods using IMU data from wearable devices effectively captured motion patterns but lacked environmental context, limiting their ability to detect anomalies resulting from interactions or surroundings. Single-modality approaches therefore suffered from reduced robustness and inconsistent performance across different operational settings.

Deep Learning Limitations for Edge Deployment: Recent advances in deep learning, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformers, have significantly improved recognition accuracy and anomaly detection [3]. However, these models typically demand high computational resources, substantial memory, and continuous power supply, which are often unavailable on edge devices such as smartphones, wearables, and embedded sensors. While techniques like model pruning and knowledge distillation can reduce model size and inference cost, their adoption remains limited, preventing efficient, real-time, on-device deployment for continuous monitoring applications.

Challenges in Multimodal Fusion: Integrating multiple modalities, such as video and IMU data, can enhance the accuracy of human activity recognition and anomaly detection, but it introduces significant challenges [4]. Different modalities often operate at varying sampling rates and produce features with distinct characteristics, making temporal alignment and joint representation learning difficult. Traditional fusion strategies, including early and late fusion, frequently fail to capture deeper relationships between modalities, limiting the potential performance gains. Moreover, multimodal systems typically demand higher computational resources and larger annotated datasets, which complicates both training and deployment on resource-constrained edge devices.

Real-Time Anomaly Detection and Edge Adaptability: Detecting anomalies in real time remains a critical challenge, particularly in sensitive applications such as healthcare monitoring and intelligent surveillance [5]. Many existing frame- works require long video sequences or extended temporal windows to make accurate

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

predictions, increasing inference time and delaying responses to urgent events. Reliance on cloud-based processing further introduces privacy concerns, network bandwidth dependency, and the risk of service interruptions. These limitations underscore the need for lightweight, multi- modal models capable of fast, accurate, and privacy-preserving anomaly detection directly on edge devices.

Anomaly Detection in Human–Robot Collaboration: In industrial and collaborative environments, accurate anomaly detection is essential for safe human–robot interaction [6]. Systems are designed to monitor human motion patterns to identify unusual or unsafe behavior that could lead to accidents. Many approaches utilize asymmetric modeling to detect rare but critical anomalies that occur infrequently in training data. Early identification of these deviations enables timely alerts, ensuring safer operations and facilitating smooth collaboration between humans and robotic systems.

Real-Time Deep Learning for Action Recognition: Deep learning techniques for real-time action recognition have gained prominence due to their ability to process live video streams efficiently [7]. These models typically utilize lightweight architectures designed to provide rapid predictions while maintaining high accuracy. Their low-latency operation makes them particularly suitable for applications such as intelligent surveillance, smart-home systems, and interactive environments, where immediate response is crucial. Optimizing these models for speed and responsiveness is essential to ensure reliable deployment on practical systems, especially when operating on devices with limited computational re- sources.

Deep Learning-Based Human Anomaly Detection: Convolutional neural networks (CNNs) have been extensively applied to detect human anomalies by learning detailed spatial patterns associated with unusual or irregular behavior [8]. These models are capable of handling challenges such as dynamic environments, imbalanced datasets, and rapidly changing scenes. By learning deep feature representations instead of relying on handcrafted features, CNN-based approaches achieve higher robustness and adaptability, making them effective for complex real-world anomaly detection tasks.

Classification-Based Anomaly Detection Using Image Features: Recent studies have leveraged image classification net- works to distinguish between normal and anomalous activities [9]. Instead of depending entirely on labeled anomaly data, these methods analyze the learned feature representations within the network to identify deviations from typical behavior. This strategy is particularly beneficial when an- notated anomaly datasets are limited or difficult to generate. Classification-based approaches integrate smoothly with existing computer vision pipelines while delivering reliable anomaly detection performance.

LSTM Autoencoder-Based Sequential Anomaly Detection: LSTM autoencoders are widely employed for detecting anomalies in sequential human activity data [10]. By learning the typical temporal patterns of normal activities, these models attempt to reconstruct them during inference, with abnormal activities producing higher reconstruction errors. Their ability to capture long-term temporal dependencies makes them suitable for both video-based and sensor-based monitoring, providing an effective mechanism for identifying deviations in human behavior over extended periods.

PROPOSED SYSTEM MODEL

This work presents a multimodal human activity recognition framework that combines video data with motion information obtained from an inertial measurement unit (IMU). Video sequences provide visual information about body posture and movement, while IMU sensors capture acceleration patterns that reflect motion dynamics. Using both sources together improves the reliability of activity recognition, particularly in situations where one modality may be affected by environmental conditions. The overall architecture of the proposed system is shown in Fig. 1.

Dataset

For this work, a multimodal dataset was collected by combining video data with sensor-based measurements obtained in real time. The visual component consists of frames captured through a webcam, while motion-related information is obtained using an accelerometer connected via an ESP- based module. Bringing these two sources together allows both appearance and movement characteristics to be captured within the same sample.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

The dataset includes six categories of activities: Driver Drowsiness, Fall or Collapse, Jogging, Sitting, Standing, and Walking, each of which contain two hundred samples. These classes were chosen to cover a mix of regular daily activities as well as situations that may indicate abnormal or critical conditions. During data collection, video frames and IMU readings were recorded together so that both modalities remain aligned over time.

Dataset/ DRIVERS_DROWSINESS/ FALL_OR_COLLAPSE/ JOGGING/

SITTING/ STAND/ WALK/

capture_YYYYMMDD_HHMMSS/ readings/

esp.txt frames/

frame_0.jpg

...

frame_29.jpg

... (200 captures for each class)

Data acquisition was performed using a simple real-time setup. The ESP device continuously transmits accelerometer values through a socket connection, while the camera captures frames in parallel using OpenCV. When recording is initiated, a new folder is created automatically using the current times- tamp, and the incoming data is stored in an organized manner. Each recording session contains a sequence of frames along with a text file that logs the corresponding sensor readings.

For every capture, around 30 consecutive frames are stored together with their matching IMU values. The dataset is arranged into separate folders for each activity class, and within each class, multiple capture sessions are maintained. Each session further contains two subdirectories: one for image frames and another for sensor data. This structure helps in keeping the dataset (well-organized) and easy to use during training and evaluation.

In summary, the dataset offers a structured combination of visual and sensor information for different human activities. It can be effectively used for developing and evaluating deep learning models focused on activity recognition as well as anomaly detection tasks.

Data Preprocessing

Prior to training, both video frames and IMU sensor readings are preprocessed to ensure consistency in model input. For the visual stream, each sample consists of 30 sequential frames. All frames are resized to a fixed resolution of 224×224 pixels to match the input requirement of the convolutional network. The pixel values are scaled to the range [0, 1] to facilitate stable training.

A light augmentation strategy is applied during data loading. In particular, horizontal flipping is randomly performed on frames with a probability of 0.5. This simple augmentation helps improve generalization without introducing excessive variation in the data.

For the IMU data, accelerometer readings along the three axes are taken from the recorded text files. The continuous data is divided into segments of 30 consecutive readings so that each segment matches the same time window as the video frames. This helps in capturing the motion pattern within a short duration. The segmented readings are then converted into numerical form and normalized so that the values are on a similar scale, which helps the model learn more effectively. Finally, each IMU segment is aligned with its corresponding video frame sequence to maintain proper synchronization between the two data sources.

Video Feature Extraction

Visual features are extracted using a convolutional neural network based on the MobileNetV2 architecture. A

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

pre-trained version of the network is employed with weights initialized from ImageNet, allowing effective feature extraction even with limited training data. To reduce overfitting, a significant portion of the earlier layers in the network is frozen during training.

Each frame in the sequence is processed independently using a Time Distributed wrapper applied to the base convolutional network. The extracted frame-level features are then passed to a Long Short-Term Memory (LSTM) layer with 32 units. This layer captures temporal dependencies across the sequence and produces a compact representation of the video segment. Dropout is applied within the LSTM to improve generalization.

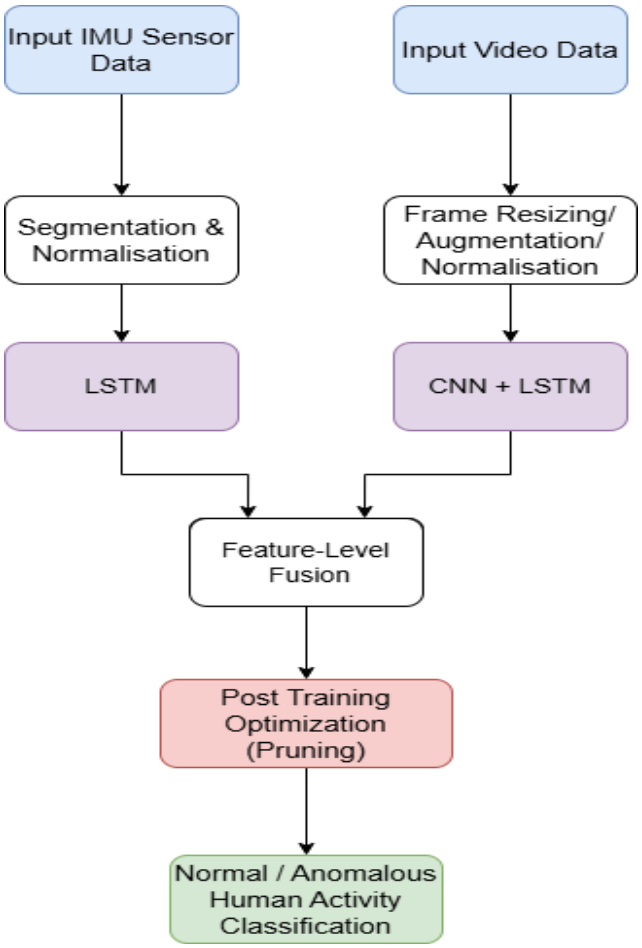

Fig. 1. Overall system design integrating video and IMU based activity recognition.

IMU Motion Modeling

The IMU data is modeled using a separate LSTM network designed to capture temporal motion patterns. The input con- sists of sequences of three-dimensional accelerometer readings over time. A smaller LSTM with 16 units is used for this branch, as the dimensionality of the sensor data is relatively low.

Dropout and recurrent dropout are incorporated to reduce overfitting and stabilize training. This branch enables the model to learn motion-related characteristics that may not be clearly visible in the visual stream alone.

Multimodal Feature Fusion

The outputs from the video and IMU branches are combined using feature-level fusion. Specifically, the feature vectors generated by the two LSTM networks are concatenated to form a unified representation.

This combined feature vector is passed through a fully connected layer with 64 units and ReLU activation. A dropout layer with a relatively high rate is applied to further reduce overfitting. Finally, a Softmax layer produces class probabilities across the six activity categories. This fusion strategy allows the model to leverage complementary information from both modalities, improving the overall robustness of activity recognition.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

Training Strategy

The model is trained using the Adam optimizer with a low learning rate of 3×10^{-5} and gradient clipping to ensure stable updates. Sparse categorical cross-entropy is used as the loss function.

The dataset is divided into training and validation sets using an 80:20 split. A generator-based approach is used to efficiently load data in batches during training. Early stopping is applied to prevent overfitting, while a learning rate scheduler reduces the learning rate when validation performance plateaus.

Post-Training Optimization using Pruning

After training, model optimization is performed using weight pruning. This technique removes less significant weights from convolutional and dense layers, introducing sparsity into the network.

A constant sparsity level of 50% is applied during pruning. The model is then restructured by removing pruning-specific operations, resulting in a compact version suitable for deployment. The optimized model requires less memory and computational resources while maintaining comparable performance.

This step makes the proposed system more practical for real-time applications, especially in environments with limited hardware capabilities.

RESULTS

The performance of the proposed multimodal CNN-LSTM model was evaluated using both training metrics and a separate test dataset. The model was trained for 50 epochs, and its learning behavior was analyzed using accuracy and loss curves.

Training Performance

Figure 2 and Figure 3 illustrate the training and validation accuracy and loss over epochs.

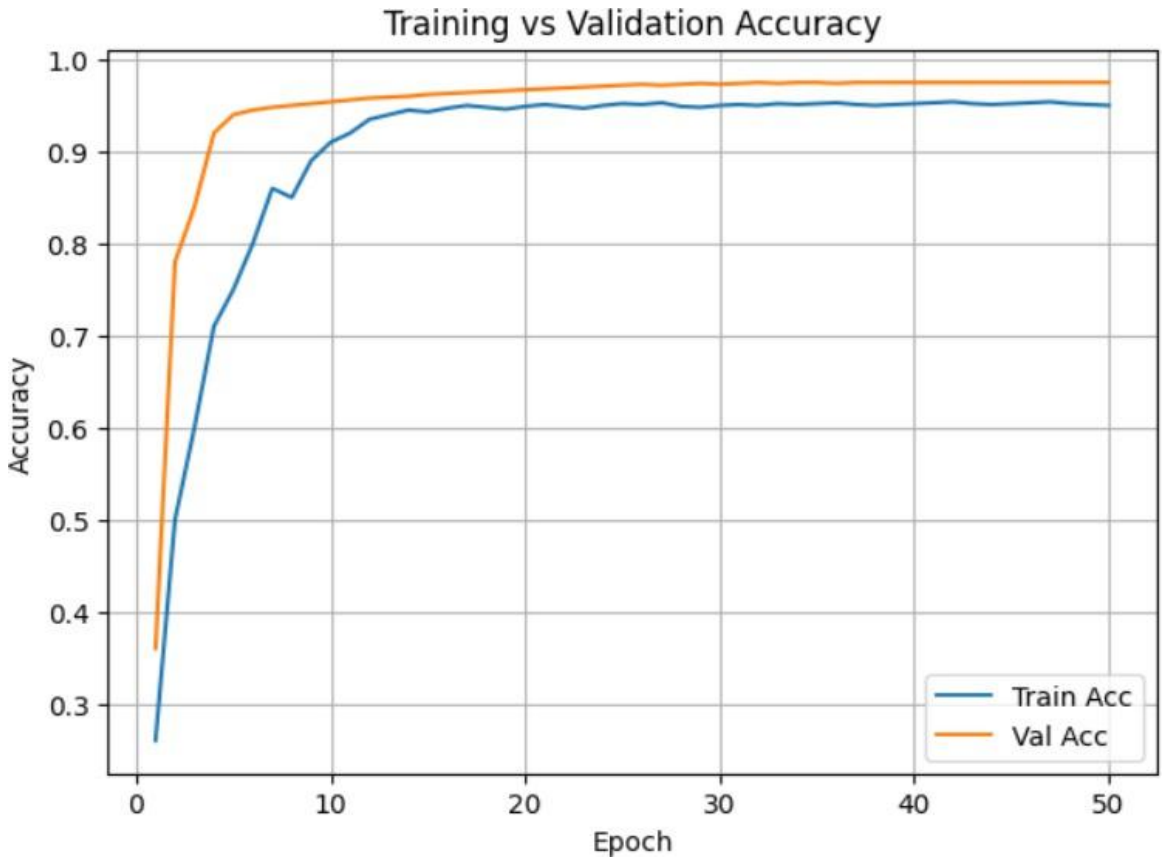

Fig. 2. Training vs Validation Accuracy

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

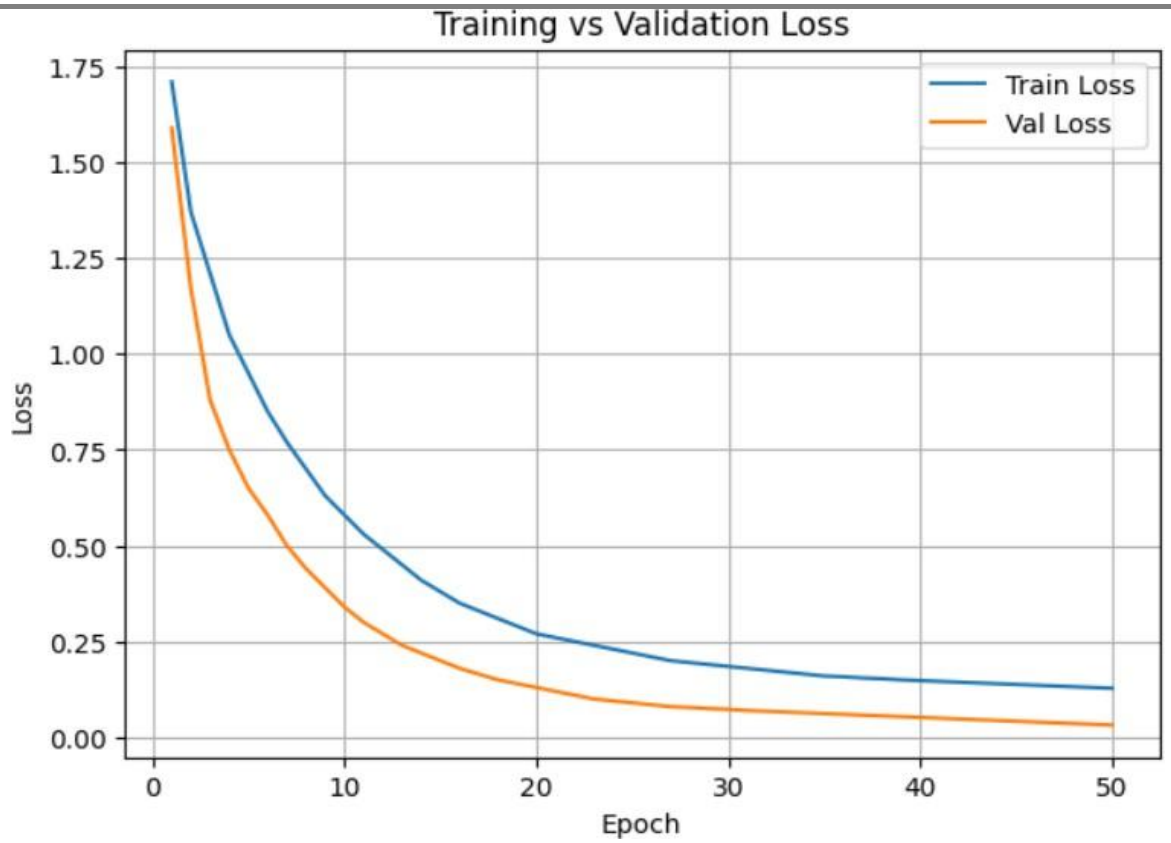

Fig. 3. Training vs Validation Loss

From the accuracy curve, it can be observed that the model achieves rapid improvement during the initial epochs and gradually stabilizes, reaching an accuracy of approximately 93-96%. Similarly, the loss curve shows a consistent decrease, indicating effective optimization and convergence without significant overfitting.

Test Evaluation

The trained model was evaluated on a separate test dataset consisting of 30 samples. The distribution of test samples across the six activity classes is shown below:

- Driver Drowsiness: 5
- Fall or Collapse: 5
- Jogging: 5
- Sitting: 5
- Standing: 5
- Walking: 5

The evaluation results are summarized as follows:

- Test Loss: 0.3069
- Test Accuracy: 92.85%

Real-Time Prediction Analysis

To assess real-world applicability, a real-time prediction interface was developed. This interface enables users to select recorded capture sessions and obtain activity predictions along with associated confidence levels.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

For regular activities, the system accurately classified nor- mal activity such as walking with a confidence level, indicating reliable recognition of common human actions. The prediction was appropriately categorized as a normal activity, confirming the effectiveness of the model in handling non-critical scenarios.

In contrast, during abnormal situations, the system success- fully identified events such as fall or collapse with confidence levels. These cases were correctly flagged as anomalies, demonstrating the model's sensitivity to high-risk conditions. The reliable performance observed in both normal and abnormal cases highlights the strength of combining visual and sensor-based data. Additionally, providing confidence scores alongside predictions enhances transparency and helps in assessing the certainty of the model's decisions.

Overall, the proposed system exhibits strong performance in both controlled testing and real-time usage, making it a suitable candidate for applications including healthcare monitoring, smart cameras, and smart surveillance solutions.

DISCUSSION

The experimental findings indicate that the proposed multimodal framework is effective in learning both visual and motion-related representations. The achieved test accuracy reflects the model's ability to generalize well when exposed to previously unseen samples. Furthermore, the gradual alignment of training and validation curves suggests a stable learning process without significant overfitting.

The model also demonstrates consistent performance across all defined activity categories. This indicates its capability to differentiate between routine human behaviors and critical abnormal events such as falls and driver drowsiness.

CHALLENGES

Despite the advantages of combining multiple data sources for activity recognition, several practical difficulties need to be addressed. Another challenge lies in the collection of high-quality multimodal datasets. Acquiring synchronized video and sensor data requires a controlled setup and careful annotation of activities. This makes the dataset creation process time-consuming and limits the availability of large-scale, diverse datasets. Consequently, models trained on such data may struggle to perform well in unseen environments or with different users. Variations in real-world conditions further complicate the task. Factors such as changes in lighting, camera positioning, and differences in sensor placement can alter the characteristics of the input data. These variations may reduce recognition accuracy unless the model is trained with sufficiently diverse samples or enhanced with adaptive mechanisms.

Aspect	**Existing Multimodal Systems**	**Proposed System**
Data Source	Public or limited datasets	Custom Video + IMU dataset
Anomaly Types	Mainly fall detection	Fall, drowsiness, daily activities
Dataset Diversity	Controlled environments	Real-time, varied conditions
Model Design	Heavy CNN/LSTM/3D models	Lightweight CNN + LSTM
Fusion Strategy	Early/Late fusion	Efficient feature-level fusion
Detection Method	Reconstruction or hybrid	Direct multi-class classification
Real-Time Capability	Limited, high latency	Yes, low-latency system
Edge Deployment	Rarely supported	Optimized with pruning
Accuracy	High but less stable	High (92–96%), stable
Application Scope	Single domain	Healthcare, driver safety, surveillance

TABLE I COMPARISON WITH EXISTING MULTIMODAL HAR SYSTEMS

FUTURE WORK

Future improvements can focus on enhancing both the efficiency and flexibility of the proposed system. One important direction is optimizing the model for deployment on resource- limited devices such as smartphones

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

or embedded systems. This can be achieved through model compression techniques, including pruning and the design of lightweight architectures. Another potential extension involves improving the way information from different modalities is combined. More sophisticated fusion strategies, such as attention-based mechanisms, could allow the model to dynamically prioritize the most relevant features from either the video or IMU data.

Increasing the size and diversity of the dataset is also essential for improving generalization. Including more participants, varied environments, and additional activity types would make the model more robust. Furthermore, integrating confidence estimation or dedicated anomaly detection modules could enhance the system's reliability in real-time applications.

CONCLUSION

This study introduces a multimodal framework for human activity recognition that leverages both visual data and motion sensor information. Spatial features are extracted from video frames using a MobileNetV2-based network, while temporal dependencies are captured using LSTM layers. Simultaneously, sequential IMU data is processed through a parallel LSTM network to model motion patterns. The outputs from both streams are then combined to perform final activity classification.

The integration of video and sensor data enables the system to utilize complementary information from each modality. While visual input provides contextual understanding of body posture and environment, IMU signals contribute precise motion-related details. This combined representation leads to improved recognition performance compared to approaches relying on a single data source.

Although the proposed method achieves strong results, there is scope for further refinement. Improving robustness under varying real-world conditions and reducing computational complexity remain important goals. Continued research in multimodal learning and efficient model design will support the development of practical and scalable activity recognition systems for applications such as healthcare monitoring, safety assistance, and intelligent environments.

REFERENCES

1. Shyma Zaidi, Jagadeesh B, Sudheesh K V, Audre Arlene A." Video Anomaly Detection and Human Activity Classification for Recognition." IEEE Conference on Computational Intelligence and Communication Networks (CICN), 2017.
2. Zhenyu Shi, J. Andrew Zhang, R. Xu, and G. Fang." Human Activity Recognition Using Deep Learning Networks with Enhanced Channel State Information." IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 2, pp. 257–265, 2018.
3. Aikaterini Tsanousa, Symeon Vrochidis, Georgios Meditskos, Ioannis Kompa Tsiaris." A Weighted Late Fusion Framework for Recognizing Human Activity from Wearable Sensors." IEEE Access, vol. 7, pp. 80673–80684, 2019.
4. D. Miki, S. Chen, and K. Demachi." Unnatural Human Motion Detection Using Weakly Supervised DNN." IEEE Access, vol. 8, pp. 194998–195010, 2020.
5. J. Yan, F. Angelini, and S. M. Naqvi." Privacy-Preserving Human Action Recognition for Anomaly Detection." IEEE Transactions on Image Processing, vol. 29, pp. 8174–8188, 2020.
6. H. Lv, P. Yi, R. Liu, Y. Hou, D. Zhou, Q. Zhang, and X. Wei." Asymmetric Anomaly Detection for Human-Robot Interaction." IEEE Transactions on Industrial Informatics, vol. 17, no. 9, pp. 6152–6162, 2021.
7. H. E. Azzag, I. E. Zeroual, and A. Ladjailia." Real-Time Human Action Recognition Using Deep Learning." IEEE Access, vol. 10, pp. 22983–22992, 2022.
8. L. Yashaswi, S. Mekala, and M. D. Prasad." Human Anomaly Detection Using Deep Learning." IEEE Xplore, 2023.
9. H.-J. Jeon, S. Lang, C. Vogel, and R. Behrens." Anomaly Detection from Image Classification." IEEE Access, vol. 12, pp. 4571–4584, 2024.S.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

10. A. Roseline, S. Karthik, and I. N. V. D. Sruti.” Intelligent Human Anomaly Detection Using LSTM Autoencoders.” IEEE Conference on Emerging Trends in Artificial Intelligence and Data Science (ETAIDS), 2024.