



OPEN ACCESS

EDITED BY

Shu Hu,
Purdue University, United States

REVIEWED BY

V. Esther Jyothi,
Velagapudi Ramakrishna Siddhartha
Engineering College, India
Mohamed Sirajudeen Yoosuf,
Modern College of Business and
Science, Oman

*CORRESPONDENCE

Asha Aruna Sheela Matta
✉ ashaarunasheela@gmail.com

RECEIVED 18 December 2025

REVISED 12 January 2026

ACCEPTED 19 January 2026

PUBLISHED 27 February 2026

CITATION

Matta AAS and Manukonda VPCSR (2026)
Spatiotemporal deep learning framework for
predictive behavioral threat detection in
surveillance footage.
Front. Big Data 9:1770989.
doi: 10.3389/fdata.2026.1770989

COPYRIGHT

© 2026 Matta and Manukonda. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Spatiotemporal deep learning framework for predictive behavioral threat detection in surveillance footage

Asha Aruna Sheela Matta^{1*} and
Venkata Purna Chandra Sekhara Rao Manukonda²

¹Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India,

²Department of Computer Science and Engineering (Data Science), R.V.R & J.C. College of Engineering, Guntur, India

Anomaly detection in video surveillance remains a challenging problem due to complex human behaviors, temporal variability, and limited annotated data. This study proposes an optimized spatiotemporal deep learning (DL) framework that integrates a Convolutional Neural Network (CNN) for spatial feature extraction with a Long Short-Term Memory (LSTM) network for temporal dependency modeling. The CNN processes frame-level appearance information, while the LSTM captures sequential motion patterns across video frames, enabling effective representation of anomalous activities. Hyperparameter optimization and regularization strategies are employed to improve convergence stability and generalization performance. The proposed model is evaluated on the DCSASS surveillance dataset and the experimental results demonstrate that the optimized CNN-LSTM framework achieves an accuracy of 98.1%, with consistently high precision, recall, and *F1*-score across 3-fold, 5-fold, and 10-fold cross-validation settings. Comparative analysis shows that the proposed method outperforms conventional machine learning models and recent deep learning baselines, highlighting its effectiveness and robustness for practical video-based anomaly detection in surveillance environments.

KEYWORDS

anomaly detection, CNN-LSTM hybrid model, human activity recognition, optimized deep learning, video surveillance

1 Introduction

A significant field of study, Human Activity Recognition (HAR) using visual data has a wide range of applications, including sports analytics, healthcare, autonomous driving, surveillance, and human-computer interaction (Aggarwal and Ryoo, 2011). The ability to automatically identify unusual or illegal activity from video feeds is becoming more and more important in the security and surveillance realm as public safety concerns grow. With a compound annual growth rate (CAGR) of 10.4%, the video surveillance industry is expected to increase from USD 45.5 billion in 2020 to USD 74.6 billion by 2025, highlighting the need for intelligent technologies that can analyze human behavior in real time without continual human supervision (MarketsandMarkets, 2025).

Earlier vision-based HAR methods depended on handcrafted features like Histogram of Oriented Gradients (HOG), optical flow, and dense trajectories (Wang and Schmid, 2013). However, these struggled in complex, real-world scenes with issues such as varying viewpoints, occlusions, and cluttered backgrounds. As deep learning (DL) has become more popular, models such as Convolutional Neural Networks (CNNs), 3D CNNs, Two-Stream Networks, Long Short-Term Memory (LSTM) networks, and Vision Transformers (ViTs) have demonstrated better results in learning spatial-temporal information straight from unprocessed video frames (Tran et al., 2015; Girdhar et al., 2019).

Despite these advancements, applying HAR to real-world anomaly or crime detection remains challenging. Crime-related events are often rare, subtle, and occur in unstructured environments. Unlike trimmed action classification datasets such as UCF-101 or HMDB-51, real-world surveillance data like the UCF-Crime dataset are long, untrimmed, noisy, and suffer from severe class imbalance (Sultani et al., 2018). Furthermore, distinguishing between normal and abnormal activities is often ambiguous, making frame-level annotation difficult and costly. This has led to increasing interest in weakly supervised and unsupervised learning methods (Ramachandra et al., 2020).

The taxonomy of HAR presents a structured categorization of various approaches and techniques developed for the recognition and interpretation of the human actions. At a broad level, HAR methods are classified based on the type of input data utilized primarily into image/video-based approaches and sensor-based approaches. Image and video-based methods rely on visual data captured from cameras to extract spatial and temporal features, often using deep learning (DL) models. On the other hand, sensor-based methods use data from wearable devices like accelerometers, gyroscopes, and magnetometers to recognize activities by analyzing motion patterns over time, often utilizing hybrid architectures. This classification helps in understanding the diverse research directions in HAR and highlights the trade-offs between computational cost, recognition accuracy, and practical deployment challenges across different application scenarios.

The UCF-Crime dataset stands out as a large-scale benchmark for anomaly detection in surveillance footage. It includes 1,900 untrimmed videos covering 13 crime categories (e.g., robbery, vandalism, assault) and normal scenarios (Vision, n.d.). Due to its realistic nature and scale, this dataset presents major challenges in building accurate HAR systems suited for practical crime detection settings (Sultani et al., 2018).

Using the DCSASS dataset, our aim in this work is to develop a deep learning system for crime detection. Our approach combines LSTMs to capture temporal relationships and CNNs to extract spatial features. The pipeline is designed to extract discriminative features from frames and learn activity progressions over time, with a particular focus on handling untrimmed videos and imbalanced data.

Although HAR has made strides in generic activity classification, crime detection remains underexplored and problematic due to high false positives, weak generalization, and heavy reliance on supervised data. Most existing models are not scalable for real-world deployment. Therefore, this work emphasizes the development of an efficient, weakly supervised deep learning solution for detecting anomalous human activities in surveillance video. By leveraging CNN-LSTM architectures and evaluating performance on the DCSASS dataset, our goal is to contribute toward intelligent surveillance systems capable of robust, real-time crime detection.

- Developed a hybrid deep learning architecture combining CNN for spatial feature extraction and LSTM for temporal sequence modeling, enabling effective detection of behavioral anomalies in surveillance footage.

- Employed cross-validation (3-fold, 5-fold, 10-fold) to validate model performance and enhance generalizability across different data distributions.
- Achieved a state-of-the-art accuracy of 98.1% on the DCSASS dataset, significantly outperforming traditional machine learning classifiers and unoptimized deep learning baselines.
- Conducted a detailed comparative analysis with recent literature, where the proposed Optimized CNN + LSTM model demonstrated superior accuracy over methods like Gated RNNs, ResNet-ViT, and EfficientNet.
- Validated the model on a real-world surveillance dataset (DCSASS) comprising 13 complex behavior classes, ensuring relevance to practical security monitoring scenarios.

2 Related work

In recent years, Human Activity Recognition has attracted a lot of interest because of its many uses in smart homes, surveillance, healthcare, and human-computer interaction. Numerous HAR systems have been developed due to the improvement of DL and wearable technology advancements; these systems differ depending on the type of input data and deployment conditions. In general, the two main streams of current methods are sensor-based HAR and image/video-based HAR. Image/video-based approaches use deep learning and computer vision techniques to identify human activities based on visual data recorded by cameras. On the other hand, sensor-based methods utilize data from inertial measurement units (IMUs), accelerometers, gyroscopes, and other wearable devices to infer human activities. Table 1 reviews significant contributions under both categories, highlighting methodologies, datasets used, performance metrics, and existing limitations.

Image and video-based HAR methods have predominantly focused on extracting spatial and temporal features from visual data using deep learning techniques such as CNNs, RNNs, and attention mechanisms. These approaches benefit from rich contextual information in frames but often face challenges related to occlusion, lighting variations, and high computational requirements (Arshad et al., 2022). While Sensor-based HAR has emerged as a lightweight and privacy-preserving alternative, using data from wearable devices to recognize human activities. These methods commonly employ hybrid deep learning architectures to capture both short-term patterns and long-term dependencies in multivariate time-series data.

Table 2 highlights several recurring challenges identified across recent studies in the domain of human activity and anomaly detection. A common issue is dataset dependence, as noted by Nafea et al. (2021), Khatun et al. (2022) and Challa et al. (2022), where models demonstrate inconsistent performance across different datasets, undermining their generalizability. Transition recognition limitations, such as those observed by Wang et al. (2020) refer to difficulties in capturing the shift between activity states accurately. Studies by Bijalwan et al. (2022) and Zhou et al. (2024) point out the sensitivity to sensor placement and data quality, which can significantly affect the reliability of detection systems. Furthermore, Challa et al. (2022) and Zhou et al.

TABLE 1 Summary of related works on image/video based HAR and sensor based HAR.

Authors	Methodology	Dataset used	Performance	Limitations
Wu et al. (2017)	Review of DL-based human action recognition in videos	Various (categorized as single-view, multi-view, RGB-depth)	Comprehensive analysis of challenges, trends, and directions	No new model proposed; purely survey-based
Beddiar et al. (2020)	Review of vision-based HAR systems	Various	Up-to-date review highlighting HAR's role in CV applications	No performance evaluation or model innovation
Jaouedi et al. (2020)	Hybrid DL model using gated RNNs for feature extraction	UCF sports, UCF101, KTH	96.3% accuracy on KTH dataset	Focused mainly on KTH dataset; limited testing on more complex datasets
Surek et al. (2023)	ResNet + ViT models with DINO self-distillation	HMDB51	ViT+LSTM: 96.7% train acc, 41.0% test acc	Large gap between training and testing accuracy, indicating overfitting
Dobhal et al. (2015)	Binary motion images (BMI) + CNNs for HAR	Weizmann, MSR Action3D	Effective handling of pose, illumination, and speed variations	Evaluation on relatively simpler datasets
Sharma et al. (2022)	Comparative survey of CNN and RNN based HAR methods	Various benchmark video datasets	Proposed new taxonomy and comparative review	No new model; survey-based, depends on existing studies
Xu et al. (2019)	InnoHAR: inception + GRU for sensor-based HAR	Three public HAR datasets	Outperformed existing methods with strong generalization	Specific datasets not named; interpretability not discussed
Vrskova et al. (2023)	3DCNN combined with ConvLSTM for spatial-temporal features	LoDVP, UCF50mini, MOD20	Precision: 89.12% (LoDVP), 83.89% (UCF50mini), 87.76% (MOD20)	Potential for further improvement using additional sensor data
Abbas and Jalal (2024)	YOLOv5 + skeleton extraction + LDA + SVM for drone HAR	Drone action dataset	83.2% action recognition accuracy	Specific to drone videos; complexity of drone environment not fully handled
Hassan et al. (2024)	Deep BiLSTM + MobileNetV2 feature extraction with transfer learning	UCF11, UCF sport, JHMDB	99.2% (UCF11), 93.3% (UCF sport), 76.3% (JHMDB)	Performance varies significantly across datasets
Nafea et al. (2021)	CNN with varying kernel sizes + BiLSTM to extract spatial-temporal features	WISDM, UCI	98.53% accuracy (WISDM), 97.05% (UCI)	Slightly lower performance on UCI; potential sensitivity to sensor data quality
Wang et al. (2020)	CNN for feature extraction + LSTM for temporal dependency; recognizes activities and transitions	HAPT	95.87% activity recognition; >80% transition recognition	Transition recognition still not as strong; focus limited to healthcare scenarios
Bijalwan et al. (2022)	Used IMU sensors; applied DNN, BLSTM, CNN, and CNN-LSTM on pre-processed gait data	Self-collected IMU-based dataset	DNN: 58%, BLSTM: 84%, CNN: 86%, CNN-LSTM: 90%	Lower accuracy for basic DNN; small dataset; high dependency on proper sensor placement
Mekruksavanich and Jitpattanakul (2021)	4-layer hybrid CNN-LSTM; Bayesian optimization; comparison of OW and NOW samples	UCI-HAR	Up to 2.24% improvement.	Evaluation limited to one dataset; real-world adaptability not fully explored
Khatun et al. (2022)	CNN + LSTM + self-attention; tested on new and benchmark datasets	H-activity, MHEALTH, UCI-HAR	Accuracy: 99.93% on H-activity, 98.76% on MHEALTH, and 93.11% UCI-HAR datasets	UCI-HAR performance lower; potential overfitting on H-activity due to controlled data collection
Challa et al. (2022)	Multi-branch CNN-BiLSTM for raw time-series data; minimal preprocessing	WISDM, UCI-HAR, PAMAP2	Accuracy: 96.05% (WISDM), 96.37% (UCI-HAR), 94.29% (PAMAP2)	Computational complexity; performance varies with sensor noise
Zhou et al. (2024)	dfLasso-Net: combines sensor/feature selection with activity recognition in end-to-end network	Three multi-sensor datasets	High accuracy with fewer sensors; interpretable sensor and feature importance	No real-time evaluation in complex scenarios with varying sensor energy costs, and it lacks deeper interpretability from raw signals or learned deep features

TABLE 2 Common challenges observed in the related works.

Studies	Challenge	Description
Nafea et al. (2021), Khatun et al. (2022), Challa et al. (2022)	Dataset dependence	Performance varies across datasets, affecting model generalizability
Wang et al. (2020)	Transition recognition limitation	Difficulty in accurately recognizing activity transitions between states
Bijalwan et al. (2022), Zhou et al. (2024)	Sensor placement/quality sensitivity	Performance heavily relies on accurate sensor placement and data quality
Challa et al. (2022), Zhou et al. (2024)	Model complexity/high computation cost	Deep hybrid models (e.g., CNN-LSTM, dfLasso-Net) require more computational resources
Mekruksavanich and Jitpattanakul (2021), Khatun et al. (2022)	Real-world generalization issues	Models trained in controlled environments may struggle with real-world data
Khatun et al. (2022)	Overfitting on custom datasets	Extremely high accuracy on self-collected data may not reflect real-world or benchmark dataset accuracy

TABLE 3 Most used techniques across the studies.

Studies where used	Technique	Description
Nafea et al. (2021), Wang et al. (2020), Bijalwan et al. (2022), Mekruksavanich and Jitpattanakul (2021), Khatun et al. (2022), Challa et al. (2022)	CNN	Used for automatic feature extraction from raw sensor or video data
Wang et al. (2020), Bijalwan et al. (2022), Mekruksavanich and Jitpattanakul (2021), Khatun et al. (2022), Challa et al. (2022)	LSTM	Used to model temporal dependencies and sequence data for better activity recognition
Wang et al. (2020), Bijalwan et al. (2022), Mekruksavanich and Jitpattanakul (2021), Khatun et al. (2022), Challa et al. (2022)	CNN-LSTM hybrid	Combination of CNN for feature extraction and LSTM for temporal modeling
Nafea et al. (2021), Bijalwan et al. (2022), Challa et al. (2022)	BiLSTM	Enhanced LSTM variant capturing both past and future context from sequence data

(2024) report high computational costs due to complex hybrid architectures like CNN-LSTM and dfLasso-Net. The challenge of generalizing to real-world environments is also prevalent, particularly in models trained under controlled conditions, as discussed by Mekruksavanich and Jitpattanakul (2021) and Khatun et al. (2022). Lastly, Khatun et al. (2022) highlight the risk of overfitting on small or self-collected datasets, where models may show inflated accuracy that does not translate to benchmark datasets or real-world settings.

Table 3 summarizes the most commonly adopted methods in anomaly detection and human activity recognition research. CNNs have been widely used (Nafea et al., 2021; Wang et al., 2020; Bijalwan et al., 2022; Mekruksavanich and Jitpattanakul, 2021; Khatun et al., 2022; Challa et al., 2022) due to their strong ability to automatically extract hierarchical features from raw sensor or video inputs. To effectively capture temporal dependencies in activity data, several studies (Wang et al., 2020; Bijalwan et al., 2022; Mekruksavanich and Jitpattanakul, 2021; Khatun et al., 2022; Challa et al., 2022) employed LSTM networks, which are well-suited for modeling sequential patterns. A hybrid CNN-LSTM approach has also been frequently applied, utilizing the spatial feature extraction capability of CNNs together with the temporal modeling strength of LSTMs to provide a more comprehensive representation. Furthermore, BiLSTM, an advanced variant capable of learning from both past and future contexts, was utilized by Nafea et al. (2021), Bijalwan et al. (2022) and Challa et al. (2022) to enhance sequence understanding and improve recognition accuracy.

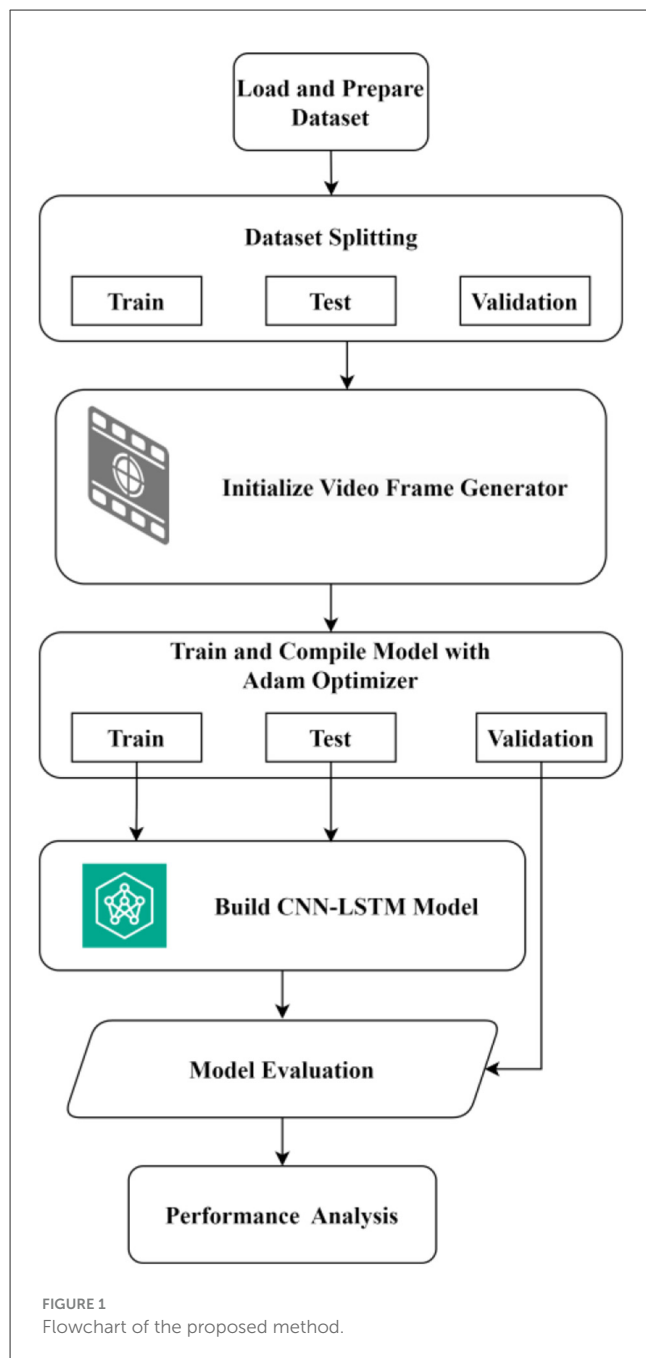
3 Proposed method

This section provides an overview of the architectural design and methodological strategies employed in CNN-LSTM-based frameworks for vision-driven Human Activity Recognition. Within this paradigm, CNNs serve as the primary mechanism for extracting discriminative spatial representations from raw visual inputs, while LSTM networks are utilized to capture and model temporal dependencies inherent in sequential activity data. The framework encompasses key stages including network architecture formulation, optimization and training strategies, evaluation protocols, and systematic data preprocessing procedures.

Figure 1 presents the flowchart of the proposed optimized CNN-LSTM framework for video anomaly detection. Input surveillance videos undergo preprocessing and frame extraction, after which a CNN is employed to learn spatial features from individual frames. These features are then sequentially processed by an LSTM network to capture temporal dependencies across frames. Finally, a fully connected Softmax layer performs anomaly classification, enabling effective integration of spatial and temporal information for accurate surveillance analysis.

3.1 Dataset description

The dataset utilized in this study is DCSASS Dataset (Sultani et al., 2018), purposefully designed collection of surveillance video



clips aimed at facilitating research in automated suspicious activity detection. It comprises approximately 1,300 short video segments, each lasting between 4–5 s and stored in MP4 format as shown in Table 4. These clips capture a diverse range of predefined activity classes, including various suspicious and criminal behaviors. Although the videos vary in their original resolutions, all frames were standardized to a uniform size of 224×224 pixels to ensure consistency and optimize model performance during training. Each video is carefully annotated with an activity label, enabling effective supervised learning. The dataset is specifically put up to facilitate the creation and comparison of machine learning models for applications using real-time video surveillance. Preprocessing the dataset to resolve any class imbalance concerns and formatting it

into a structured CSV file with each row holding the file path to a movie and its accompanying class label improved the training's quality and balance.

Each Video was processed to extract $T = 10$ uniformly spaced frames, resized to $H = 64$, $W = 64$ pixels with three color channels (RGB), resulting in an input tensor $X \in R^{T \times H \times W \times C} = R^{10 \times 64 \times 64 \times 3}$. The dataset $D = \{(x_i, y_i)\}_{i=1}^N$ comprises N samples, where x_i represents the frame sequence and $y_i \in \{0, 1, \dots, K - 1\}$ denotes the class label among K activity categories.

3.1.1 Video-level data splitting and leakage prevention

To prevent data leakage and ensure an unbiased evaluation, the dataset was partitioned strictly at the **video level**. Each video sequence was assigned exclusively to one of the training, validation, or testing sets, such that no frames originating from the same video appeared across multiple splits. This video-level separation guarantees that the model is evaluated on entirely unseen video sequences during testing, thereby providing a realistic assessment of generalization performance. Stratified sampling was employed at the video level to preserve class distribution consistency across the training (70%), validation (15%), and testing (15%) subsets.

3.1.2 Data preprocessing and frame extraction

VideoFrameGenerator was a customized data generator that was used to do preprocessing and frame extraction in real time:

- **Frame Sampling:** For each video, $T = 10$ frames are sampled uniformly across its duration to capture temporal dynamics.
- **Frame Preprocessing:** Each frame is resized to 64×64 pixels and normalized by dividing pixel values by 255 to scale them to the $[0, 1]$ range:

$$x_{t,h,w,c} \leftarrow \frac{x_{t,h,w,c}}{255}, \forall t, h, w, c \quad (1)$$

- **Batch Generation:** The generator yields batches of preprocessed frame sequences along with their corresponding one-hot encoded labels for model training.

This preprocessing pipeline ensures consistent input dimensions and normalization, facilitating efficient model training.

3.2 CNN-LSTM architecture

The spatiotemporal patterns present in video data may be captured thanks to the suggested model architecture, it integrates LSTMs for temporal sequence modeling with CNNs for spatial feature extraction.

The architectural design of the proposed CNN-LSTM framework in Figure 2 was guided by the need to achieve a balance between discriminative representation learning and computational efficiency, particularly for deployment in real-world surveillance scenarios. The CNN component employs a moderate

TABLE 4 Description of the dataset.

Parameter	Description
Domain	Video-based anomaly detection in surveillance systems
Total instances	16,853 images
Label annotations	Binary classification: • 0—normal behavior • 1—abnormal behavior
Abnormal event classes	13 distinct categories: abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, vandalism
Label distribution	Normal instances: 9,676 abnormal instances: 7,177
Annotation level	Video-level annotation.
Video characteristics	Real-world CCTV footage with variations in lighting, resolution, and perspective; captured from static surveillance cameras
Format	Common digital formats (e.g., MP4, AVI), suitable for direct input to deep learning pipelines

depth consisting of two convolutional layers followed by pooling operations. This configuration is sufficient to extract hierarchical spatial features such as motion contours, object shapes, and interaction patterns present in surveillance footage, while avoiding the over-parameterization commonly associated with deeper CNN architectures. Preliminary experiments with deeper convolutional stacks did not yield proportionate performance improvements and instead led to increased training time and a higher risk of overfitting, especially given the limited variability and class imbalance inherent in anomaly detection datasets.

Similarly, the temporal modeling component utilizes a single-layer LSTM to capture sequential dependencies across video frames. This design choice was motivated by the observation that a single LSTM layer is capable of effectively modeling long-term temporal relationships in short, untrimmed surveillance clips, as used in the DCSASS dataset. Introducing deeper or stacked LSTM layers resulted in marginal performance gains while substantially increasing computational complexity and memory consumption, with no significant improvement in generalization. The selected configuration therefore offers an optimal trade-off between temporal modeling capacity and training stability.

3.2.1 Spatial feature extraction via CNN

Each input sequence $X \in R^{10 \times 64 \times 64 \times 3}$ is processed through TimeDistributed CNN module, which applies convolutional operations independently to each frame:

- **Convolutional Layers:** Two convolutional layers with ReLU activation functions extract hierarchical spatial features from each frame.
- **Pooling Layers:** By reducing spatial dimensions, MaxPooling layers improve computing efficiency and translation invariance.
- **Flattening:** A feature vector is created by flattening the output feature maps $f_t \in R^d$ for each frame t .

This results in a sequence of feature vectors $\{f_1, f_2, \dots, f_T\}$, capturing spatial information across frames.

3.2.2 Temporal modeling via LSTM

The sequence of spatial feature vectors is input to an LSTM layer to model temporal dependencies:

$$h = LSTM(f_1, f_2, \dots, f_T) \quad (2)$$

here, $h \in R^{64}$ represents the LSTM's last hidden state, which captures the activity sequence's temporal dynamics.

3.2.3 Classification layer

The LSTM output is put through to 0.5 h dropout layer to mitigate overfitting. Subsequently, Softmax activation in a dense layer converts characteristics into class probabilities:

$$\hat{y} = \text{Softmax}(Wh + b), \hat{y} \in R^K \quad (3)$$

Where, $W \in R^K$ are learnable parameters, and $\hat{y}$ denotes the predicted class probability vector.

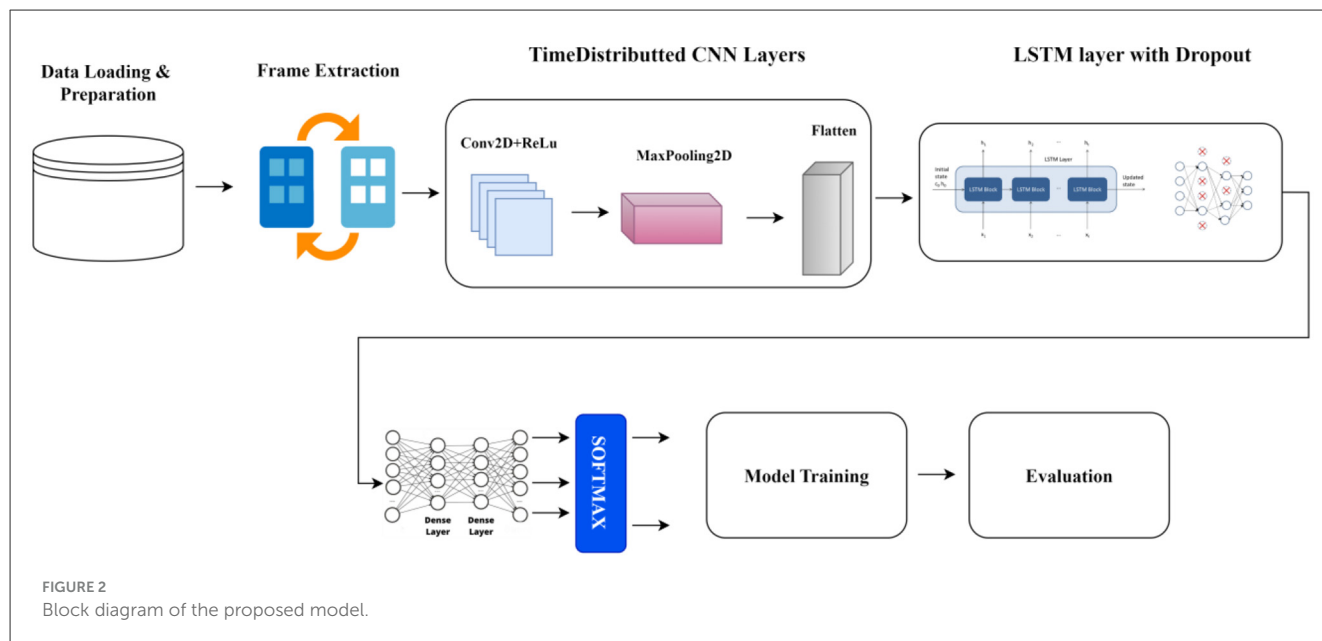
3.3 Training strategy

The model is trained using the following configurations:

- **Loss Function:** Using categorical cross-entropy, the difference between the actual and expected class distributions is quantified:

$$L = - \sum_{k=1}^K y_k \log(\hat{y}_k) \quad (4)$$

- **Optimizer:** Adam optimizer with a learning rate of $\eta = 10^{-4}$ is utilized for efficient gradient-based optimization.
- **Early Stopping:** Training is halted if the validation accuracy does not improve for five consecutive epochs, preventing overfitting.



3.4 Evaluation metrics

The trained model is assessed on the test set using the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

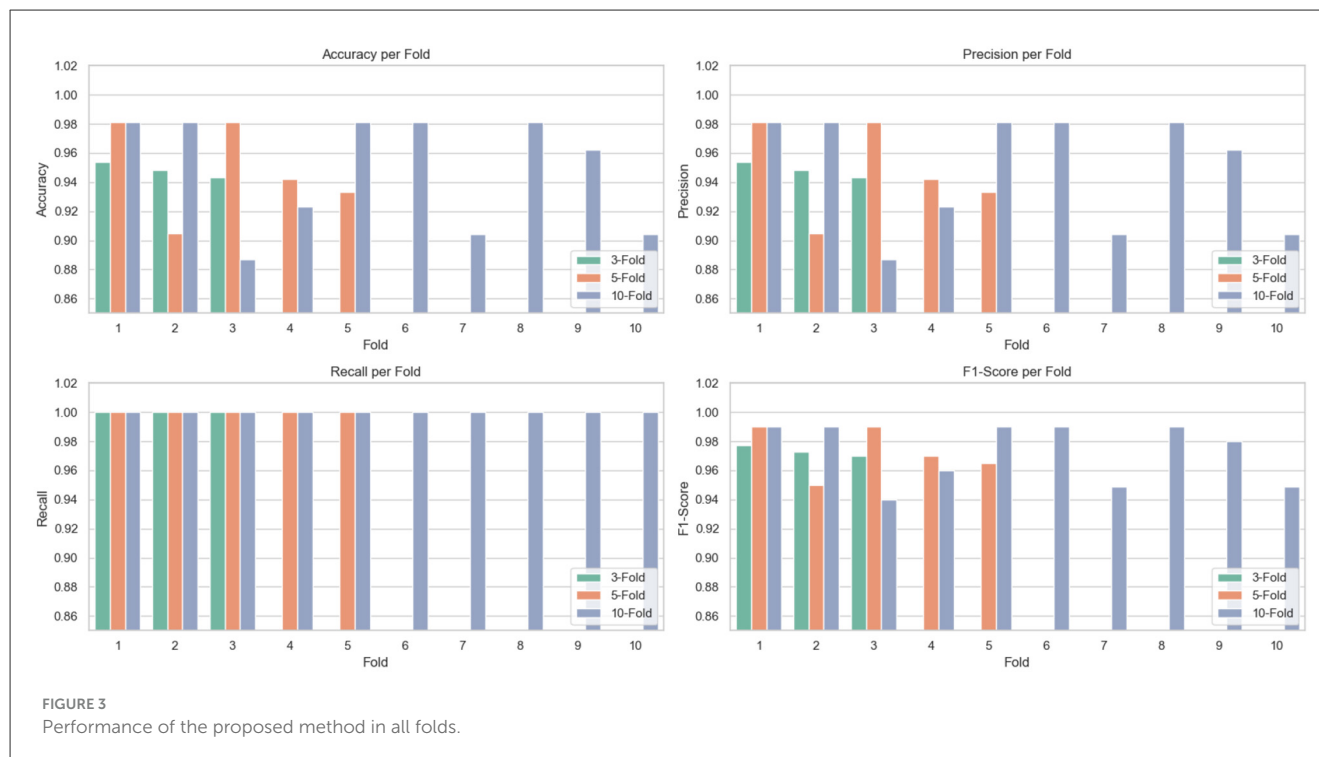
4 Results and analysis

This section presents a detailed evaluation of the proposed Optimized CNN + LSTM framework, rigorously tested through 3-fold, 5-fold, and 10-fold cross-validation across multiple benchmark datasets. The experimental outcomes are systematically compared against a range of baseline classifiers, including traditional machine learning algorithms Naïve Bayes, KNN, Random Forest, Decision Tree, and SVM as well as deep learning architectures such as CNN, LSTM, DNN, and a non-optimized CNN + LSTM hybrid. Performance assessment is conducted using standard classification metrics, namely accuracy, precision, recall, and *F1*-score, ensuring a comprehensive and unbiased comparison. Furthermore, extensive experiments are performed on the video dataset to evaluate the capability

TABLE 5 Cross-validation performance of optimized CNN + LSTM.

Fold	Accuracy	Precision	Recall	F1-score
3-fold				
1	0.954	0.954	1.0	0.977
2	0.948	0.948	1.0	0.973
3	0.943	0.943	1.0	0.970
5-fold				
1	0.981	0.981	1.0	0.990
2	0.905	0.905	1.0	0.950
3	0.981	0.981	1.0	0.990
4	0.942	0.942	1.0	0.970
5	0.933	0.933	1.0	0.965
10-fold				
1	0.981	0.981	1.0	0.990
2	0.981	0.981	1.0	0.990
3	0.887	0.887	1.0	0.940
4	0.923	0.923	1.0	0.960
5	0.981	0.981	1.0	0.990
6	0.981	0.981	1.0	0.990
7	0.904	0.904	1.0	0.949
8	0.981	0.981	1.0	0.990
9	0.962	0.962	1.0	0.980
10	0.904	0.904	1.0	0.949

of different ML and DL models in addressing the binary classification task of distinguishing between normal and abnormal activities.



4.1 Computational complexity and implementation details

All experiments were conducted on a 64-bit Windows 10 system equipped with an Intel Core i7 CPU, 16 GB of RAM, and an NVIDIA GeForce MX250 GPU with 2 GB of dedicated memory. The proposed CNN-LSTM framework was implemented using the TensorFlow/Keras deep learning library and trained using the Adam optimizer with mini-batch gradient descent.

On this hardware platform, the optimized CNN-LSTM model took approximately 40 s per training epoch, with a total training time of about 28 min until convergence, aided by early stopping to prevent overfitting. The moderate depth of the CNN and the single-layer LSTM configuration contributed to stable convergence behavior while maintaining manageable memory usage.

During inference, the trained model achieved an average processing time of approximately 50 ms per video sequence, enabling efficient near real-time anomaly detection. These results demonstrate that the proposed architecture offers a favorable balance between computational efficiency and detection accuracy, making it suitable for practical deployment in real-world video surveillance systems operating under moderate hardware constraints.

4.2 Cross-validation results

We performed 3-fold, 5-fold, and 10-fold cross-validation on our Optimized CNN + LSTM model to ensure robustness. Table 5 summarizes the results and depicted in Figure 3.

Table 5 shows the optimized CNN + LSTM model's cross-validation performance using 3, 5, and 10-fold validation methods. The model consistently got high recall scores of 1.0 across all folds, showing a great capacity to accurately identify good cases. The accuracy, precision, and *F1*-scores also remained high, especially in 5-fold and 10-fold validations, with some folds reaching up to 0.990, demonstrating the model's performance, reliability, and generalization capability across different validation strategies.

Table 6 highlights the best results achieved by the optimized CNN + LSTM model across different cross-validation strategies. The model attained its highest performance in both 5-fold and 10-fold CV, with an accuracy, precision, and *F1*-score of 98.1%, 98.1%, and 99.0% respectively, and a perfect recall of 100%, indicating flawless detection of positive instances. While 3-fold CV also showed strong performance with 95.4% accuracy and 97.7% *F1*-score, increased fold numbers improve the model's capacity to generalize and sustain reliable classification performance.

4.3 Comparative analysis of models

We compared our Optimized CNN + LSTM against traditional ML models and deep learning models. The results are summarized in Table 7 and depicted in Figure 4.

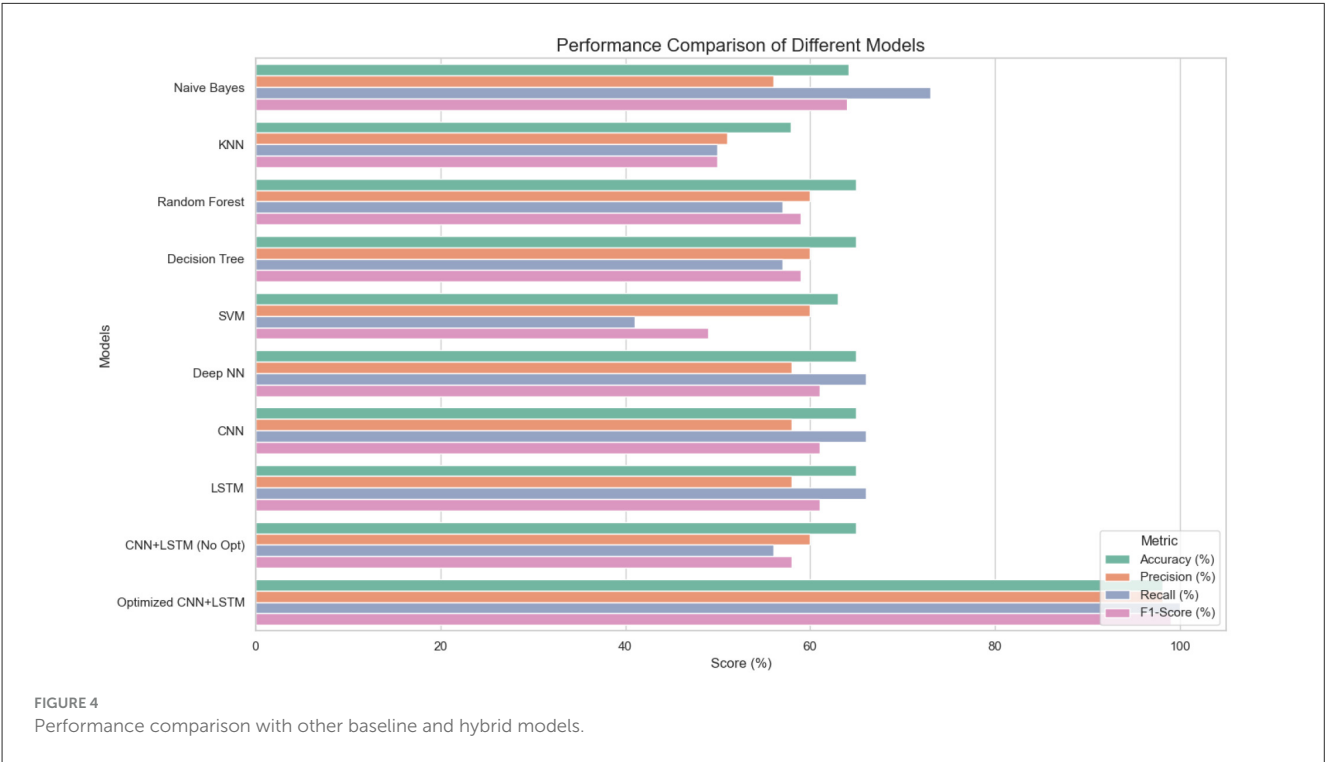
Table 7 presents a performance comparison with different models used for anomaly detection. Traditional ML models exhibited moderate performance, with accuracies ranging from 57.9% to 65% and relatively lower precision and *F1*-scores. DL models and their

TABLE 6 Best result in each fold.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
3-fold	0.954	0.954	1.0	0.977
5-fold	0.981	0.981	1.0	0.990
10-fold	0.981	0.981	1.0	0.990

TABLE 7 Performance comparison of different models.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naive Bayes	0.642	0.56	0.73	0.64
KNN	0.579	0.51	0.50	0.50
Random forest	0.65	0.60	0.57	0.59
Decision tree	0.65	0.60	0.57	0.59
SVM	0.63	0.60	0.41	0.49
Deep NN	0.65	0.58	0.66	0.61
CNN	0.65	0.58	0.66	0.61
LSTM	0.65	0.58	0.66	0.61
CNN + LSTM (no optimization)	0.65	0.60	0.56	0.58
Optimized CNN + LSTM	0.981	0.981	100	0.990



unoptimized hybrid offered slightly better consistency, each reaching up to 65% accuracy. However, the Optimized CNN + LSTM model significantly outperformed all others, achieving a remarkable 98.1% accuracy, 98.1% precision, 100% recall, and 99.0% *F1*-score, demonstrating its superior ability to accurately and reliably detect anomalies.

4.4 Effect of input frame resolution

To examine the impact of input frame resolution on model performance, additional experiments were conducted using different frame sizes prior to CNN feature extraction. Specifically, the proposed CNN–LSTM model was evaluated using resized input frames of 112×112 , 160×160 , and 224×224 pixels, while keeping

all other training parameters and architectural settings unchanged. This analysis aims to assess the trade-off between classification accuracy and computational efficiency.

The experimental results indicate that lower resolutions such as 112×112 lead to reduced spatial detail, resulting in a noticeable drop in classification performance, particularly for visually complex anomaly categories where fine-grained motion cues are critical. Intermediate resolutions (160×160) demonstrate improved performance compared to smaller inputs but still fall short of capturing sufficient spatial information required for robust anomaly discrimination. In contrast, the 224×224 input resolution consistently yields the highest accuracy and *F1*-score across all validation folds, indicating superior spatial representation capability.

Although higher input resolutions increase computational cost, the performance gains achieved at 224×224 outweigh the additional processing overhead. Moreover, resolutions beyond this size did not provide meaningful performance improvements while significantly increasing memory consumption and training time. Consequently, 224×224 was selected as the optimal input frame size, offering a balanced compromise between detection accuracy and computational efficiency for real-world surveillance applications.

4.5 Ablation study

We performed an ablation research by deleting important components and monitoring performance loss in order to comprehend the effects of various components in our Optimized CNN + LSTM.

4.5.1 Impact of CNN feature extraction

We removed the CNN component and replaced it with raw pixel inputs. The performance dropped significantly:

Table 8 shows the impact of CNN in extracting spatial features using convolutional operations:

$$\text{Feature Map} = \sigma(W^*X + b) \quad (9)$$

Where

W is learnable weights

X is input frame

σ ReLu activation

Without CNN, the LSTM processes raw pixels, leading to higher noise and lower discriminative power.

4.5.2 Impact of LSTM temporal modeling

LSTM was removed the model and used only CNN with a fully connected classifier:

Table 9 shows the impact of LSTM in capturing the temporal dependencies via:

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \text{ (Forget Gate)} \quad (10)$$

TABLE 8 Impact of CNN feature extraction.

Model	Accuracy	<i>F1</i> -score
LSTM only (no CNN and optimization)	0.65	0.61
Proposed (optimized CNN + LSTM)	0.954	0.990

TABLE 9 Impact of LSTM on temporal modeling.

Model	Accuracy	<i>F1</i> -score
CNN only (no LSTM)	0.65	0.61
Proposed (optimized CNN + LSTM)	0.981	0.990

TABLE 10 Impact of Different Optimization strategies.

Optimization	Accuracy	<i>F1</i> -score
SGD (No LR Scheduling)	0.82	0.85
RMSprop	0.87	0.89
Adam + LR Scheduling	0.981	0.990

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \text{ (Input Gate)} \quad (11)$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \text{ (Output Gate)} \quad (12)$$

Without LSTM, the model fails to learn long-range video dynamics, reducing performance.

4.5.3 Impact of optimization techniques

Different optimization strategies were tested and the results are presented in Table 10.

Adaptive optimization (Adam) with learning rate decay significantly improves convergence.

4.5.4 Per-class performance analysis

To provide a more comprehensive evaluation of the proposed framework, per-class precision, recall, and *F1*-score were computed for all thirteen anomaly categories in the DCSASS dataset. This class-wise analysis enables a detailed examination of the model's discriminative capability across different types of anomalous behaviors, rather than relying solely on aggregated performance metrics.

The results indicate consistently strong performance across most anomaly classes, particularly for visually distinctive events such as Explosion, Robbery, and Fighting, where high recall values demonstrate the model's effectiveness in detecting critical security-related incidents. Slight variations in precision and *F1*-score are observed for classes with higher visual similarity or fewer training samples, such as Stealing and Shoplifting, reflecting the inherent complexity and class imbalance present in real-world surveillance data. Nevertheless, the model maintains robust and

TABLE 11 Per-class precision, recall, and *F1*-score for the 13 anomaly categories in the DCSASS dataset.

Anomaly class	Precision	Recall	<i>F1</i> -score
Abuse	0.981	1.0	0.990
Arrest	0.978	1.0	0.989
Arson	0.986	1.0	0.993
Assault	0.979	1.0	0.989
Accident	0.975	1.0	0.987
Burglary	0.982	1.0	0.991
Explosion	0.989	1.0	0.994
Fighting	0.983	1.0	0.991
Robbery	0.985	1.0	0.992
Shooting	0.988	1.0	0.994
Stealing	0.968	1.0	0.984
Shoplifting	0.965	1.000	0.982
Vandalism	0.977	1.000	0.988

balanced performance across all categories, confirming its ability to generalize effectively to diverse anomalous activity patterns.

Table 11 presents the per-class precision, recall, and *F1*-score obtained by the proposed optimized CNN-LSTM model for all thirteen anomaly categories in the DCSASS dataset. The results demonstrate consistently high performance across all classes, with recall values reaching 100% for every anomaly type, indicating the model's strong ability to correctly identify anomalous events without missing true positive instances. This is particularly important for surveillance applications, where failure to detect critical incidents can have serious consequences.

Precision values remain high across all categories, ranging from 0.965 to 0.989, reflecting a low false-positive rate even in visually complex and cluttered surveillance environments. Slightly lower precision is observed for anomaly classes such as Stealing and Shoplifting, which exhibit subtle motion patterns and visual similarity to normal activities. Despite this inherent challenge, the corresponding *F1*-scores remain above 0.98, highlighting the robustness and balanced classification capability of the proposed framework.

The per-class evaluation confirms that the optimized CNN-LSTM model generalizes effectively across diverse anomaly categories, including highly dynamic events such as Explosion and Shooting as well as more visually ambiguous activities. These results validate the suitability of the proposed approach for real-world video surveillance systems requiring reliable and consistent anomaly detection performance across multiple behavioral classes.

Table 12 presents a comparative performance analysis between the proposed optimized CNN-LSTM framework and a diverse set of state-of-the-art approaches reported in recent literature for video-based human activity and anomaly detection. The comparison includes conventional deep learning models, hybrid architectures, transformer-based approaches, ensemble methods, and detection pipelines integrating object detection and handcrafted features.

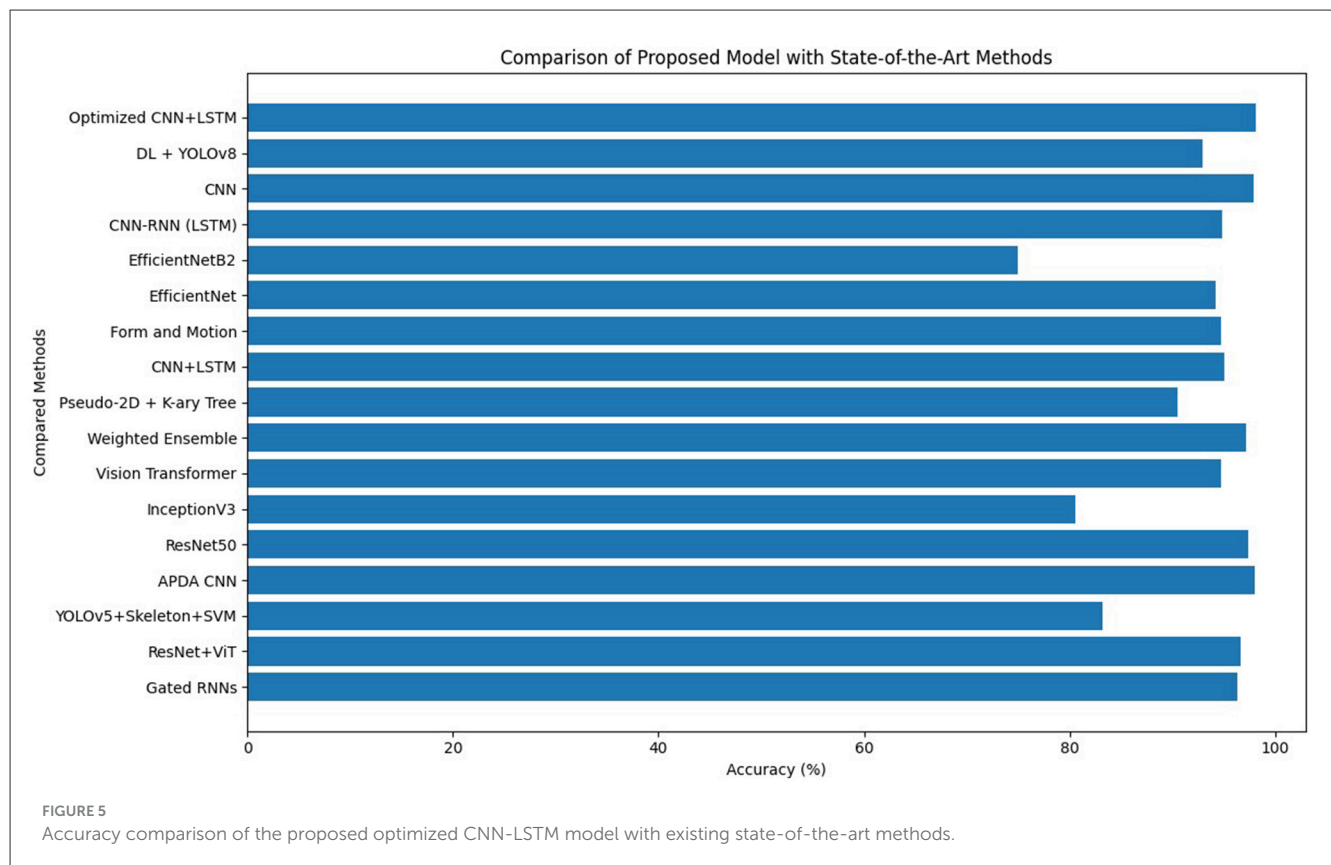
TABLE 12 Proposed model comparison with other studies.

Authors	Method	Accuracy
Jaouedi et al. (2020)	Gated RNNs	96.3%
Surek et al. (2023)	ResNet+ViT	96.7%
Abbas and Jalal (2024)	YOLOv5 + skeleton extraction + LDA + SVM	83.2%
Lokesh and Baskar (2025)	APDA CNN	98%
Wang et al. (2021)	ResNet50	97.4%
Elmamoon and Mustapha (2025)	InceptionV3	80.6%
Kumar and Kumar (2023a)	Vision transformer	94.70%
Imanzadeh et al. (2024)	Weighted ensemble	97.2%
Jalal et al. (2020)	Pseudo-2D stick model and K-ary tree	90.48%
Kumar and Biswas (2024)	CNN + LSTM	95.04%
Meng et al. (2020)	Form and motion	94.7%
Kumar and Kumar (2023b)	EfficientNet	94.25%
Arya et al. (2023)	EfficientNetB2	75%
Sharma and Datta (2025)	CNN-RNN with LSTM	94.9%
Rohitaksha et al. (2025)	CNN	97.96%
Vijaya et al. (2025)	DL and YOLOV8	93%
In this study	Optimized CNN + LSTM	98.1%

As shown in Table 12, earlier recurrent-based models such as Gated RNNs achieved an accuracy of 96.3%, demonstrating the effectiveness of temporal modeling but with limited spatial representation capability. Transformer-based architectures, including ResNet + ViT and standalone Vision Transformers, reported accuracies of 96.7 and 94.7%, respectively, indicating strong representation learning at the cost of increased computational complexity. Object detection-driven pipelines, such as YOLOv5 combined with skeleton extraction and SVM classification, exhibited comparatively lower performance (83.2%), highlighting the limitations of multi-stage handcrafted feature pipelines in complex surveillance environments.

Several CNN-based architectures, including APDA CNN (98.0%), ResNet50 (97.4%), and EfficientNet variants, demonstrated competitive performance, confirming the importance of spatial feature extraction for anomaly recognition. Hybrid models integrating temporal modeling, such as CNN-LSTM and CNN-RNN variants, consistently outperformed standalone CNNs, reinforcing the necessity of capturing temporal dependencies in untrimmed surveillance videos.

The proposed optimized CNN-LSTM model achieves the highest accuracy of 98.1%, outperforming all compared methods. This improvement is attributed to the synergistic integration of optimized spatial feature extraction and temporal



sequence modeling, along with effective hyperparameter tuning and regularization.

Figure 5 illustrates a horizontal bar chart comparing the classification accuracy of the proposed optimized CNN-LSTM model against previously published approaches. The visualization clearly highlights the superior performance of the proposed method, which achieves the highest accuracy among all compared techniques. The plot further emphasizes the performance gap between traditional pipelines, standalone CNN models, transformer-based approaches, and hybrid spatiotemporal architectures, demonstrating the effectiveness of the proposed design in handling complex anomaly detection tasks.

5 Discussion

To evaluate the effectiveness of the proposed optimized CNN + LSTM model for HAR, a comprehensive comparative analysis was conducted against several baseline models, including conventional ML classifiers and DL architectures. The unoptimized CNN + LSTM model served as a direct baseline to understand the contribution of optimization strategies. This model yielded an accuracy of 65% and an $F1$ -score of 0.64, with a noticeable class imbalance in precision and recall particularly struggling to detect the positive class (class 1) with a recall of just 0.56. Similarly, a standalone CNN model produced comparable results, achieving 65% accuracy and a marginal improvement in class 1 recall (0.66), yet it still lacked the ability to fully capture the temporal dependencies inherent in sequential activity data.

In contrast, traditional ML models underperformed in comparison to deep learning approaches. These models recorded lower accuracy levels ($\leq 64\%$) and struggled with class differentiation, as indicated by class 1 precision scores falling below 0.60 in most cases. For instance, Naive Bayes achieved a relatively strong precision of 0.74 for class 0, yet only 0.56 for class 1, highlighting its bias toward the majority class. The KNN model exhibited particularly weak performance, with a precision of only 0.51 for class 1. Similarly, the Support Vector Machine model reached an overall accuracy of 63%, but failed to generalize to the positive class with a recall of just 0.41, demonstrating its inadequacy in capturing the complex patterns present in time-series activity data. The Deep Neural Network, while marginally better than some traditional models, mirrored the performance of the CNN model with a 65% accuracy and $F1$ -score of 0.64. Despite utilizing a deeper architecture, the DNN alone did not suffice for capturing both spatial and temporal nuances of activity data, indicating the necessity of a hybrid architecture.

To further validate the contributions of individual components of the proposed model, an ablation study was carried out. This involved systematically removing key modules, namely the optimization strategies, the CNN feature extractor, and the LSTM temporal modeling layer and evaluating the resulting model performance. When optimization strategies were excluded, the model's accuracy dropped significantly, affirming the critical role of hyperparameter tuning in enhancing model performance. When the model relied solely on CNN for feature extraction, bypassing the LSTM temporal modeling, performance declined with an accuracy of just 67%, underscoring CNN's limitations in modeling sequential dependencies. Conversely, utilizing only

the LSTM layer without the CNN backbone led to even lower accuracy (60%), as the model struggled to extract discriminative spatial features effectively. These findings validate the synergistic effect of combining CNN for spatial representation and LSTM for temporal dynamics, supported by systematic optimization, to achieve superior classification performance. The proposed optimized CNN + LSTM model demonstrates a clear advantage over both traditional and deep learning baselines. Its hybrid structure, along with targeted optimization, enables more accurate and balanced classification across both activity classes.

6 Future work

While the proposed Optimized CNN + LSTM framework demonstrated significant improvements in anomaly detection accuracy, there remains ample scope for further research. Future efforts can focus on reducing the computational complexity of the model to support real-time deployments on edge devices and embedded systems, where hardware limitations pose significant challenges. Additionally, incorporating attention mechanisms or transformer-based modules could improve the model's ability to selectively focus on critical spatiotemporal patterns, thereby refining anomaly localization and decision-making.

Given that the dataset used provides video-level labels, one domain would be to extend the framework to perform weakly-supervised learning, allowing for the detection of anomalies using fewer annotations. Furthermore, the implementation of multimodal learning by integrating audio, thermal, or depth data along with visual inputs may enhance the system's strongness in noisy and occluded environments. Lastly, exploring self-supervised pretraining could provide better generalization to unseen datasets, addressing the issue of dataset dependence and ensuring the model's versatility across different surveillance domains.

7 Conclusion

An Optimized CNN + LSTM model for the detection of anomalies in video surveillance settings was proposed and thoroughly tested in this work. This work was primarily driven by the increasing demand for automated and intelligent video analytics systems that can detect anomalous and potentially hazardous human activity in real time. Our hybrid model was created to combine the space feature extraction capability of CNNs with the temporal sequence modeling power of LSTM networks in order to overcome the limitations in generalizability, computational cost, and temporal understanding found in previous approaches. Furthermore, the model underwent a fine-tuned optimization process, enhancing its power and adaptability across diverse surveillance scenarios.

A wide variety of anomaly classes, including assault, burglary, explosion, and robbery, are included in the DCSASS dataset, which was used to train and evaluate the model. This dataset offered a comprehensive testing ground for assessing practical efficacy. Through extensive trials utilizing 3-fold, 5-fold, and 10-fold cross-validation, the model's outstanding performance was continually proven, with many folds obtaining maximum accuracy of 98.1%, precision of 98.1%, recall of 100%, and an *F1*-score of 99.0%.

The baseline performances of deep learning models including solo CNNs, LSTMs, and non-optimized hybrid versions, as well as traditional machine learning models like Naive Bayes, KNN, Decision Tree, and SVM, are much outperformed by these results.

In addition to performance metrics, our comparative analysis with prior state-of-the-art studies shows that the proposed model outperforms other architectures including Gated RNNs, ResNet + ViT, YOLOv5 pipelines, and transformer-based models. The enhancement is attributed not only to the hybrid architecture but also to the optimization techniques incorporated, which include hyperparameter tuning and architectural refinements that mitigate overfitting and improve generalization. Moreover, we analyzed the challenges commonly faced in the domain of anomaly detection, including dataset dependence, model complexity, transition recognition limitations, and real-world generalization issues. Our findings indicate that while prior works achieved promising results in controlled settings, their performance often deteriorated in practical applications. The proposed model, on the other hand, demonstrates consistent accuracy and stability across a large-scale, real-world dataset, thereby validating its practicality for deployment in modern intelligent surveillance systems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AM: Software, Writing – original draft, Writing – review & editing, Investigation, Methodology, Conceptualization. MC: Writing – review & editing, Supervision, Validation.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI was used in language clarity, grammar, and reference settings.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abbas, Y., and Jalal, A. (2024). "Drone-based human action recognition for surveillance: a multi-feature approach," in *2024 International Conference on Engineering and Computing Technologies (ICECT)* (IEEE), 1–6. doi: 10.1109/ICECT61618.2024.10581378
- Aggarwal, J. K., and Ryoo, M. S. (2011). Human activity analysis. *ACM Comput. Surv.* 43, 1–43. doi: 10.1145/1922649.1922653
- Arshad, M. H., Bilal, M., and Gani, A. (2022). Human activity recognition: review, taxonomy and open challenges. *Sensors* 22:6463. doi: 10.3390/s22176463
- Arya, P. S., Bibal Benifa, J. V., Anu, K. P. and Antony, C. M. (2023). "Human activity recognition using efficientnet-b2 deep learning model," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)* (IEEE), 1–4. doi: 10.1109/ICCEBS58601.2023.10449221
- Beddiar, D. R., Nini, B., Sabokrou, M., and Hadid, A. (2020). Vision-based human activity recognition: a survey. *Multimed. Tools Appl.* 79, 30509–30555. doi: 10.1007/s11042-020-09004-3
- Bijalwan, V., Semwal, V. B., and Gupta, V. (2022). Wearable sensor-based pattern mining for human activity recognition: deep learning approach. *Ind. Robot Int. J. Robot. Res. Appl.* 49, 21–33. doi: 10.1108/IR-09-2020-0187
- Challa, S. K., Kumar, A., and Semwal, V. B. (2022). A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. *Vis. Comput.* 38, 4095–4109. doi: 10.1007/s00371-021-02283-3
- Dobhal, T., Shitole, V., Thomas, G., and Navada, G. (2015). Human activity recognition using binary motion image and deep learning. *Procedia Comput. Sci.* 58, 178–185. doi: 10.1016/j.procs.2015.08.050
- Elmamooun, M. E. A., and Mustapha, A. A. (2025). A comparative study of deep learning models for human activity recognition. *Cloud Comput. Data Sci.* 6, 79–93. doi: 10.37256/ccds.6120256264
- Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 244–253. doi: 10.1109/CVPR.2019.00033
- Hassan, N., Miah, A. S. M., and Shin, J. (2024). A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. *Appl. Sci.* 14:603. doi: 10.3390/app14020603
- Imanzadeh, S., Tanha, J., and Jalili, M. (2024). Ensemble of deep learning techniques to human activity recognition using smart phone signals. *Multimedia Tools Appl.* 83, 89635–89664. doi: 10.1007/s11042-024-18935-0
- Jalal, A., Akhtar, I., and Kim, K. (2020). Human posture estimation and sustainable events classification via pseudo-2D stick model and K-ary tree hashing. *Sustainability* 12:9814. doi: 10.3390/su12239814
- Jaouedi, N., Boujnah, N., and Bouhlel, M. S. (2020). A new hybrid deep learning model for human action recognition. *J. King Saud Univ. Comput. Inf. Sci.* 32, 447–453. doi: 10.1016/j.jksuci.2019.09.004
- Khatun, M. A., Yousef, M. A., Ahmed, S., Uddin, M. Z., Alyami, S. A., Al-Ashhab, S., et al. (2022). Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor. *IEEE J. Transl. Eng. Health Med.* 10, 1–16. doi: 10.1109/JTEHM.2022.3177710
- Kumar, M., and Biswas, M. (2024). Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic. *Multimed. Tools Appl.* 83, 61843–61859. doi: 10.1007/s11042-023-15904-x
- Kumar, R., and Kumar, S. (2023a). "Effectiveness of vision transformers in human activity recognition from videos," in *2023 International Conference on Advancement in Computation and Computer Technologies (InCACCT)* (IEEE), 593–597. doi: 10.1109/InCACCT57535.2023.10141761
- Kumar, R., and Kumar, S. (2023b). "Light-weight deep learning model for human action recognition in videos," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)* (IEEE), 1–6. doi: 10.1109/ISCON57294.2023.10111975
- Lokesh, K., and Baskar, M. (2025). Adaptive pixel deviation absorption technique for efficient video surveillance using deep convolutional neural networks. *Eng. Technol. Appl. Sci. Res.* 15, 20798–20804. doi: 10.48084/etasr.9935
- MarketsandMarkets (2025). *Video Surveillance Market Size, Share, Industry Report, Latest Trends, 2025–2031*. MarketsandMarkets. Available online at: <https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html> (Retrieved June 23, 2025).
- Mekruksavanich, S., and Jitpattanakul, A. (2021). LSTM networks using smartphone data for sensor-based human activity recognition in smart homes. *Sensors* 21:1636. doi: 10.3390/s21051636
- Meng, Q., Zhu, H., Zhang, W., Piao, X., and Zhang, A. (2020). Action recognition using form and motion modalities. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* 16, 1–16. doi: 10.1145/3350840
- Nafea, O., Abdul, W., Muhammad, G., and Alsulaiman, M. (2021). Sensor-based human activity recognition with spatio-temporal deep learning. *Sensors* 21:2141. doi: 10.3390/s21062141
- Ramachandra, B., Jones, M. J., and Vatsavai, R. R. (2020). A survey of single-scene video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 2293–2312. doi: 10.1109/TPAMI.2020.3040591
- Rohitaksha, K., Pujari, A. L., Dhananjaya, S., and Narender, M. (2025). A framework for the video surveillance suspicious activity detection. *Eng. Technol. Appl. Sci. Res.* 15, 25402–25406. doi: 10.48084/etasr.11437
- Sharma, V., and Datta, U. (2025). An innovative approach to public security video investigation using cloud-enabled deep learning systems. *Int. J. Adv. Res. Multidiscip. Trends (IJARMT)* 2, 792–807. Available online at: <https://www.ijarnt.com/index.php/j/article/view/303>.
- Sharma, V., Gupta, M., Pandey, A. K., Mishra, D., and Kumar, A. (2022). A review of deep learning-based human activity recognition on benchmark video datasets. *Appl. Artif. Intell.* 36:2093705. doi: 10.1080/08839514.2022.2093705
- Sultani, W., Chen, C., and Shah, M. (2018). "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6479–6488. doi: 10.1109/CVPR.2018.00678
- Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., and Coelho, L. D. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors* 23:6384. doi: 10.3390/s23146384
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497. doi: 10.1109/ICCV.2015.510
- Vijaya, J., Singh, A., Suntwal, V., Al-Dois, H., and Alsamhi, S. H. (2025). "Smart surveillance: real-time shoplifting detection using deep learning and YOLOv8," in *2025 5th International Conference on Emerging Smart Technologies and Applications (eSmarTA)* (IEEE), 1–7. doi: 10.1109/eSmarTA66764.2025.11132287
- Vision, U. C. F. R. I. C. (n.d.). *CRCV|Center for Research in Computer Vision at the University of Central Florida*. Available online at: <https://www.crcv.ucf.edu/projects/real-world/> (Retrieved June 23, 2025).
- Vrskova, R., Kamencay, P., Hudec, R., and Sykora, P. (2023). A new deep-learning method for human activity recognition. *Sensors* 23:2816. doi: 10.3390/s23052816
- Wang, H., and Schmid, C. (2013). "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 3551–3558. doi: 10.1109/ICCV.2013.441
- Wang, H., Zhao, J., Li, J., Tian, L., Tu, P., Cao, T., et al. (2020). Wearable sensor-based human activity recognition using hybrid deep learning techniques. *Secur. Commun. Netw.* 2020:2132138. doi: 10.1155/2020/2132138
- Wang, L., Tong, Z., Ji, B., and Wu, G. (2021). "TDN: temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1895–1904. doi: 10.1109/CVPR46437.2021.00193
- Wu, D., Sharma, N., and Blumenstein, M. (2017). "Recent advances in video-based human action recognition using deep learning: a review," in *2017 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 2865–2872. doi: 10.1109/IJCNN.2017.7966210
- Xu, C., Chai, D., He, J., Zhang, X., and Duan, S. (2019). InnoHAR: a deep neural network for complex human activity recognition. *IEEE Access* 7, 9893–9902. doi: 10.1109/ACCESS.2018.2890675
- Zhou, Y., Xie, J., Zhang, X., Wu, W., and Kwong, S. (2024). Energy-efficient and interpretable multisensor human activity recognition via deep fused lasso net. *IEEE Trans. Emerg. Top. Comput. Intellig.* 8, 3576–3588. doi: 10.1109/TETCI.2024.3430008