Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/IN05.20250400
Published: November 6, 2025

An Intelligent Deep Learning Approach for Detecting Suspicious Human Behavior in Surveillance Footage

Umaira Shahneen
Student, P.D.A. College of Engineering, Kalaburagi, India
umairashahneenkhan@gmail.com

Namra Mahveen and Umaima Farzeen Khan
Student, Faculty of Medicine, Khaja BandaNawaz University, Kalaburagi, India
khannamra07@gmail.com, khanumaimaf03@gmail.com

S. M. Hasanuddin
Student, Methodist College of Engineering and Technology, Hyderabad, India
s.hasanuddin20@gmail.com

Ayesha Fatima and Saheba Aijaz
Students, Stanley College of Engineering and Technology for Women
Hyderabad, India
ayeshafatimaNMEIS@gmail.com, sahebaa05@gmail.com

Qutubuddin Syed Mohammed
Professor, Industrial & Production Engineering
P.D.A. College of Engineering, Kalaburagi, India
syedqutub16@gmail.com

Abstract

The recognition of suspicious human behavior (SHAR) has become an essential component of modern surveillance and public safety systems, as it assists in identifying and preventing potential threats in real-world scenarios. This research presents an advanced framework designed to accurately detect and classify suspicious human movements using deep learning-based techniques. While several studies have explored this area, many existing approaches still face challenges such as limited detection accuracy and high computational demand. To overcome these issues, the proposed work introduces a robust methodology that combines effective data preprocessing, systematic training, and optimized deep learning models. The framework employs Convolutional Neural Networks (CNNs) along with time-distributed CNN and Conv3D architectures to achieve superior recognition performance. The experimental evaluation demonstrates accuracy rates of 90.14% and 88.23% for the respective models, outperforming conventional approaches reported in prior research. In addition, the trained models are tested on unseen data and real-time YouTube surveillance videos to validate their reliability and adaptability. These evaluations confirm that the models can generalize effectively to new situations and detect suspicious actions in practical applications. The findings highlight the potential of the proposed system to significantly enhance security and monitoring infrastructures by providing faster, more accurate, and automated detection of abnormal human activities across various environments.

© IEOM Society International

Keywords
Suspicious Human, Activity Recognition (SHAR), Deep Learning, Convolutional Neural Networks (CNN), Conv3D, Time-Distributed CNN

1. Introduction
In recent years, technology has become deeply connected to everyday life, influencing how safety and security are maintained in both public and private areas. Among various innovations, video-based monitoring systems have gained special importance for observing human actions and detecting unusual situations. Identifying activities that differ from normal human behavior has therefore become a key research problem in computer vision. With the increase in incidents across busy environments such as transport stations, financial institutions, and public facilities, it is crucial to have systems that can automatically notice and report suspicious movements. Relying only on people to watch camera footage is slow, tiring, and inefficient, which makes manual observation difficult for large-scale monitoring setups. This challenge has encouraged the design of automatic systems that can analyze live or recorded video data without constant human attention.

Artificial intelligence and deep learning techniques now make it possible to build such systems that can recognize human behavior directly from video inputs. The aim of the present work is to develop a deep learning model that can identify and label six key human actions—Running, Punching, Falling, Snatching, Kicking, and Shooting. The system is intended to support faster decision-making and improve safety measures by detecting these behaviors accurately. Convolutional Neural Networks (CNNs) are particularly suitable for this task because they can extract useful patterns and visual features from each video frame, allowing for efficient activity classification.

In addition to spatial features, learning the time-based flow of motion is also essential. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures can process sequential data, helping the system understand how an action unfolds over time. Long-term temporal convolution methods have also been used to improve recognition by learning extended patterns of movement. Although several improvements have been made in this area, challenges remain in expanding the variety of detectable activities and in maintaining accuracy under different lighting conditions, camera angles, and environmental changes. Addressing these limitations can lead to more adaptive and reliable surveillance systems that contribute to better public security.

2. Literature Survey
Ghazal et al. (2021) performed a comparative study on human activity recognition using two-dimensional skeletal data. Their research applied the OpenPose framework to extract joint-based motion and visual information from 2D landmarks. The team compared five supervised learning methods—Support Vector Machine (SVM), Naïve Bayes, Linear Discriminant, k-Nearest Neighbors (KNN), and a feed-forward backpropagation neural network—to classify four actions: sitting, standing, walking, and falling. Among these, the KNN algorithm achieved the highest recognition accuracy, highlighting its suitability for activity-based motion detection.

Zhu et al. (2020) introduced a continuous human action recognition model that utilized skeletal data captured by Kinect depth sensors. Their method relied on a variable-length Maximum Entropy Markov Model (MEMM) to recognize actions in real time without predefined start and end points. Similarly, another study explored depth-based bone information combined with machine learning to accurately classify human movements captured by depth cameras.
Hbali et al. (2019) proposed a skeleton-based unified framework to analyze the spatial and temporal dimensions of human actions. Their model measured differences between joint coordinates using Minkowski and cosine distance metrics, evaluated on public datasets such as MSR Daily Activity 3D and MSR 3D Motion. By employing the Extremely Randomized Tree algorithm, their approach demonstrated strong performance for elderly monitoring applications, particularly due to its use of low-cost sensors and open-source computational tools.

Karpathy et al. (2014) made significant contributions to video-based activity classification by applying Convolutional Neural Networks (CNNs) to a dataset of over one million videos spanning 487 categories. They developed a multi-resolution CNN architecture capable of capturing both spatial and temporal information, leading to substantial accuracy improvements—from 43.9% to 63.3%—after fine-tuning on the UCF-101 dataset. Their study established CNNs as a foundational model for video understanding.

© IEOM Society International

Feichtenhofer et al. (2016) further extended the use of CNNs by integrating both spatial and temporal data streams for action detection. Their dual-network framework combined spatial appearance and motion cues at a convolutional level, resulting in high accuracy across benchmark video datasets. The work demonstrated the benefits of merging space-time features without increasing computational cost.

Li et al. (2018) presented a CNN-based approach for indoor human activity recognition utilizing location-aware data. Their model architecture, which included convolutional, pooling, and fully connected layers, successfully classified six types of indoor activities with an accuracy of 86.7%, validating the practicality of CNNs for indoor behavioral analysis.

Anishchenko (2019) focused on fall detection using deep and transfer learning approaches. By modifying the AlexNet CNN architecture and incorporating temporal heuristics, the model achieved improved precision in identifying fall sequences captured from surveillance footage, demonstrating the potential of deep learning in health and safety monitoring.

Gul et al. (2021) applied the YOLO (You Only Look Once) network to real-time patient surveillance. By retraining the model for 32 epochs on labeled behavior images, they achieved a remarkable accuracy of 96.8% in recognizing human activities, showcasing YOLO's capability in fast and accurate action recognition.

Ullah et al. (2022) proposed an anomaly detection framework that combined pre-trained CNNs for feature extraction with a Bidirectional LSTM (BD-LSTM) for temporal analysis. Evaluated on the UCF-Crime dataset, their hybrid model delivered strong performance in identifying abnormal events within video surveillance systems, proving its effectiveness for security applications.

3. Methodology

This study proposes a systematic approach to identify suspicious human actions through video analysis. The workflow begins with gathering raw video and image data from two independent sources, denoted as S1 and S2. These sources are combined and carefully processed to correct inconsistencies, unify formats, and remove irrelevant information, resulting in a clean and integrated dataset ready for analysis.

To prepare the visual data, each image is resized and adjusted to maintain consistency across the dataset. Additional transformations, such as rotation, flipping, and color adjustments, are applied to expand the variety of training samples, which helps the models generalize better to unseen scenarios. Labels are assigned to frames showing potentially suspicious actions, providing structured data for supervised learning.

The dataset is then divided into separate segments: one for training the models and another for evaluating their predictive performance. This separation ensures that the evaluation reflects the model's ability to handle new, unseen data, rather than merely memorizing the training samples.

Feature extraction is carried out using Convolutional Neural Networks (CNNs), which analyze each frame to identify critical visual patterns. These extracted features are fed into advanced temporal models, including Hybrid LSTM, Time-Distributed CNN, Keras-GRU, and Conv3D networks. By combining spatial information with temporal sequences, these models learn to recognize complex movement patterns associated with suspicious behaviors.

Once the models are trained, they are deployed to monitor live or recorded video feeds, such as surveillance cameras or online videos. The system flags potentially unusual human activities in real time, demonstrating its usefulness in enhancing safety and security monitoring.

Overall, this approach offers a flexible and robust solution for automated activity detection. By integrating advanced feature extraction and deep learning, it improves the ability of surveillance systems to identify and respond to potential threats efficiently.

4. System Architecture

The dataset for this research was specifically developed for video classification. Since no existing dataset met the study's requirements, the researchers created their own by collecting and organizing relevant videos. The dataset

© IEOM Society International

includes six action categories: Falling, Kicking, Running, Punching, Shooting, and Snatching. These videos were compiled through systematic collection and curation to ensure quality and diversity. The videos were obtained from two online platforms, referred to as Sources (S1 and S2). These platforms were chosen for their accessibility and the availability of videos relevant to the target actions. After collection, the videos were categorized into separate folders based on their respective classes. For example, videos showing falling behavior were placed in a folder labeled "Falling," while those depicting other actions were organized similarly according to their category. Videos depicting kicking activities were placed in a folder labeled "Kicking." This systematic organization facilitated efficient management and retrieval of videos during data preparation and model training. The final dataset comprised 564 videos for training and 142 for testing. Although there were slight variations in the number of samples across categories, each class maintained a distinct set of videos. For instance, the "Snatching" class included 113 videos, while the "Kicking" class contained 119 videos. The distinct class representations emphasize the importance of maintaining a well-balanced dataset to minimize bias and enhance the model's ability to generalize effectively across different action categories (Figure 1).

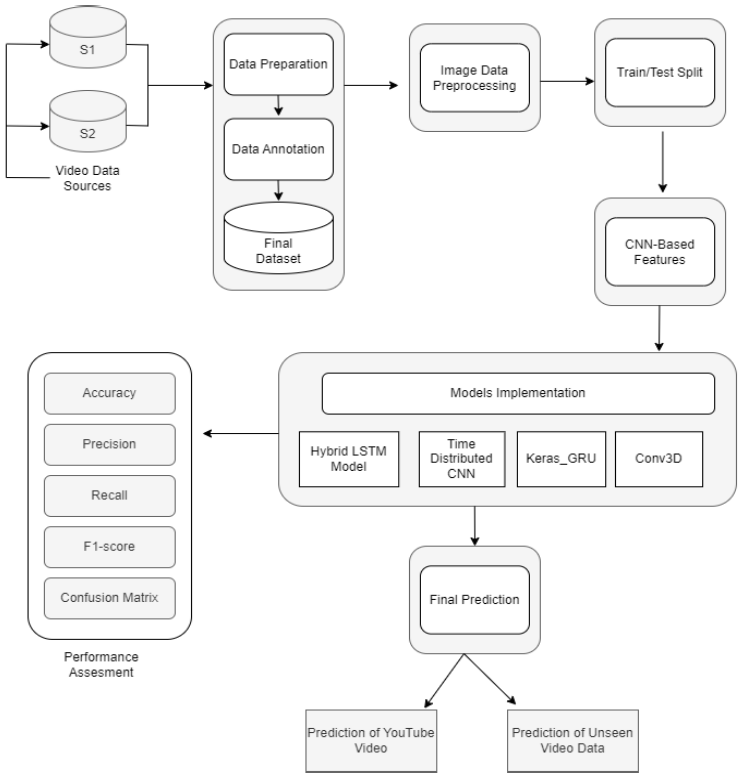


Figure 1. Approach for Suspicious Human Activity Recognition.

5. Data Collection

Data annotation was automated using Python to improve efficiency. The script scanned all video files in the specified directory, extracted their names, and stored them in both a cache and a CSV file. This process streamlined annotation by consolidating all video names into a single labeled CSV file. After annotation, all relevant files were merged into a unified label.csv, which served as the primary reference for dataset preparation and model training. The dataset was then divided into training and testing subsets to ensure independent data for model development and evaluation (Figure 2).

© IEOM Society International

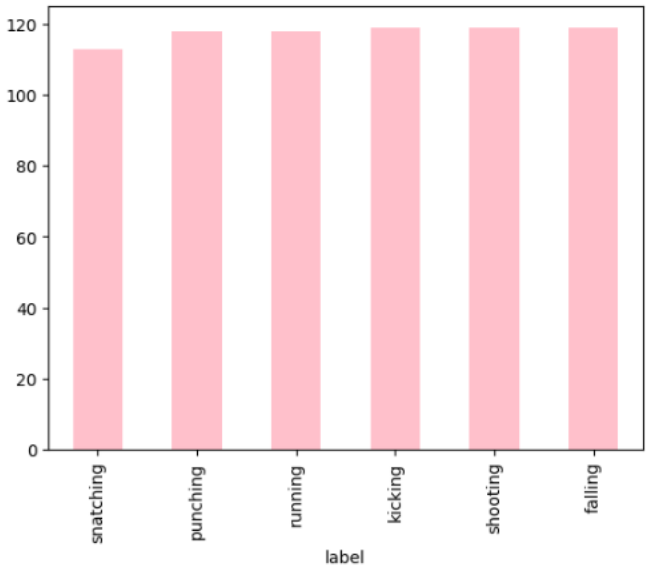


Figure 2. Dataset Distribution

6. Data Pre-Processing

Data preprocessing is a crucial step in research implementation, particularly for video analysis tasks. In this study, video data was preprocessed using the Python OpenCV library. An empty list was first created to store the extracted frames. Each video file was accessed through its corresponding label, allowing the total frame count to be determined. To extract frames systematically, a fixed interval was calculated based on a sequence length of 20 frames. This interval was obtained by dividing the total number of frames by the sequence length, ensuring uniform frame extraction across all videos.

Frames were extracted at regular intervals to ensure comprehensive coverage of each video's content. After extraction, all frames were resized to standardized dimensions of either 224×224 pixels or 64×64 pixels, depending on the analysis requirements. This resizing ensured consistency in frame size across all samples, enabling uniform processing and evaluation throughout the study.

7. Feature Extraction

Feature extraction was performed using the InceptionV3 model, a variant of CNNs known for its strong performance in image analysis. The model was customized with specific parameters, including the use of pre-trained ImageNet weights, removal of the top classification layer for feature extraction, application of average pooling for dimensionality reduction, and adjustment of the input shape to match the video frame dimensions. Using this configuration, InceptionV3 analyzed each video frame to extract 2,048 key features essential for classification. The model's architecture, captures complex visual patterns within the frames, enabling efficient feature representation for subsequent evaluation and classification.

8. Models Implementation

Detecting suspicious human activities in video recordings depends on robust deep learning models. We examine several architectures—Hybrid LSTM [21], Time Distributed TimeNN [22], Keras_GRU [23], and Conv3D [24]—each designed to capture spatio-temporal patterns in video sequences. These models use different techniques for feature extraction and sequence modeling. Through experiments, we aim to determine the optimal model for accurately detecting and classifying suspicious activities (Figure 3 and Table 1).

© IEOM Society International

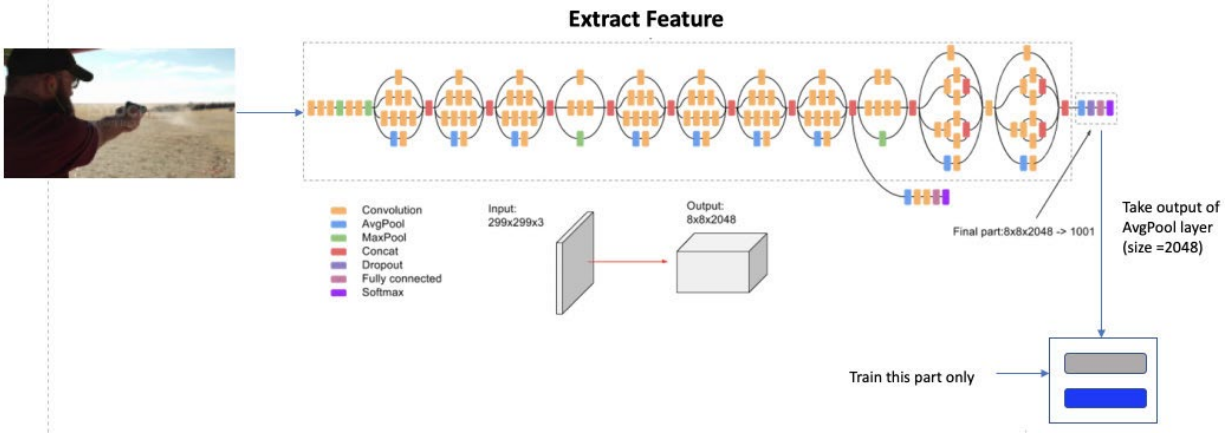


Figure 3. Features Extraction Method.

Table 1. Models Parameters.

Model	Layers	Epochs	Activation	Loss Function	Optimizer
Time Distributed CNN	Conv2D, MaxPooling2D, Dropout, LSTM, Flatten, Dense	100	Softmax	Sparse Categorical Cross entropy	Adam
Keras_GRU	GRU, Dropout, Dense	200	Softmax	Sparse Categorical Cross entropy	Adam
Hybrid Model	Conv2D, ConvLSTM MaxPooling3D TimeDistributed Flatten Dense	100	Softmax	Categorical Cross entropy	Adam
Conv3D	Conv3D, MaxPooling3D, Flatten, Dropout, Dense	20	Softmax	Sparse Categorical Cross entropy	Adam

9. Models Implementation

Detecting suspicious human activities in videos relies on robust deep learning models. We evaluate several architectures—Hybrid LSTM [21], Time Distributed TimeNN [22], Keras GRU [23], and Conv3D [24]—designed to capture spatio-temporal patterns in video sequences. Each model has unique strengths in feature extraction and sequence modeling. Through experiments, we aim to identify the best-performing model for accurately detecting and classifying suspicious activities. Detailed model parameters are given in Table 1.

10. Proposed Work Algorithm

Algorithm presents the systematic approach followed in this study. It begins with Dataset Selection, sourcing data from two origins, S1 and S2, to establish a foundation for analysis. In Data Preparation, the collected data is cleaned, structured, and integrated to ensure integrity. Next, Image Data Preprocessing applies scaling, augmentation, and normalization to improve input quality.

During Data Annotation, instances of suspicious human behavior are labeled to support supervised learning. The dataset is then split into training (80%) and testing (20%) sets to evaluate model performance on unseen data. CNN-based Feature Extraction generates informative representations from preprocessed images.
Performance Assessment uses metrics such as accuracy, precision, recall, F1-score, and confusion matrix to evaluate model efficacy. In Model Implementation, deep learning models-including Time Distributed CNN, Keras_GRU, Conv3D, and Hybrid LSTM—are trained using the extracted features. Finally, in Final Prediction, the trained models are used to detect suspicious activities in unseen videos, demonstrating practical real-world applicability.

11. Experimental Analysis and Results

During this stage, the study concentrates on analyzing how different deep learning techniques perform in recognizing and classifying suspicious human activities captured in video data. The core aim is to process individual video frames

© IEOM Society International

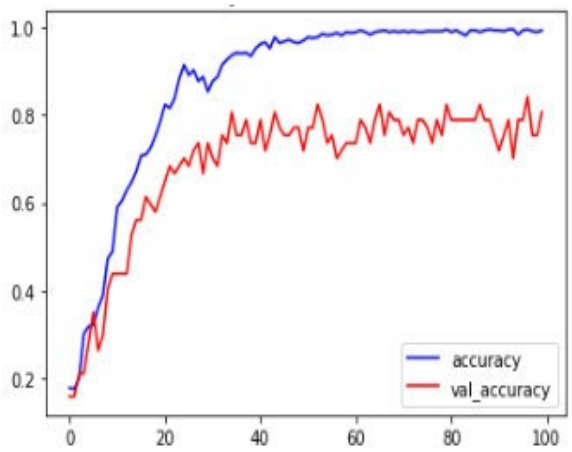
and assign them to one of six predefined behavior categories. Architectures explored for this purpose include Time DistributedNN, Keras_GRU, the Hybrid Model, and Conv3D. Each model is carefully trained, tested, and compared to determine its capability and reliability in accurately detecting abnormal behavioral patterns present in the dataset. The effectiveness of these models is measured using key evaluation metrics such as accuracy, precision, recall, F1-score, and a detailed analysis of the confusion matrix. To maintain fairness and consistency in evaluation, the dataset is divided into two distinct parts: 564 videos are utilized for training, providing sufficient learning data, while 142 videos are exclusively used for testing. This systematic separation helps ensure that the models are evaluated objectively on unseen data, preserving the integrity and validity of the experimental analysis.

11.1. Experimental Configuration

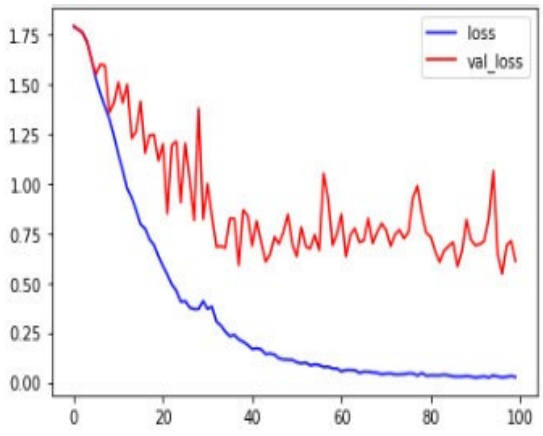
The analysis was conducted using Python version 3.8 within the Kaggle IDE environment. Since training deep learning models is often computationally intensive and time-consuming, it is essential to ensure that all required libraries are installed properly for smooth model training and execution. TensorFlow, one of the most widely adopted frameworks, plays a key role in building efficient image-processing models. The libraries employed in this study included TensorFlow, Keras, Scikit-learn, Matplotlib, Pillow, and OpenCV. Keras serves as a high-level API integrated with TensorFlow, simplifying the creation and management of deep learning architectures. Scikit-learn, a versatile Python package, was used to implement machine learning algorithms such as classification and regression. Matplotlib supported the visualization of data and results, while OpenCV and Pillow were applied for image manipulation and preprocessing tasks. Proper installation and configuration of these libraries were essential to ensure the seamless execution of all experimental procedures.

11.1.1 Hybrid Model

The Hybrid Model achieved an accuracy of 84.51%, with corresponding precision, recall, and F1-score values of 84.89%, 83.10%, and 84.72%, respectively. This architecture combines the strengths of CNN and LSTM networks to effectively capture both spatial and temporal features within the video data. As shown in Figure 4(a), the training phase demonstrates a steady improvement in accuracy over successive epochs, reflecting the model's increasing ability to learn meaningful patterns from the data. However, the validation accuracy tends to plateau around the 25th epoch, implying that the model has reached its optimal performance threshold. Meanwhile, the loss curve illustrated in Figure 4(b) shows a consistent decline, signifying enhanced predictive accuracy. This downward trend in both training and validation loss indicates that the model generalizes well and does not exhibit overfitting. Furthermore, the confusion matrix in Figure 4(c) provides deeper insights into class-wise performance, highlighting specific areas where the model may encounter difficulty in achieving precise classification (Figure 4).

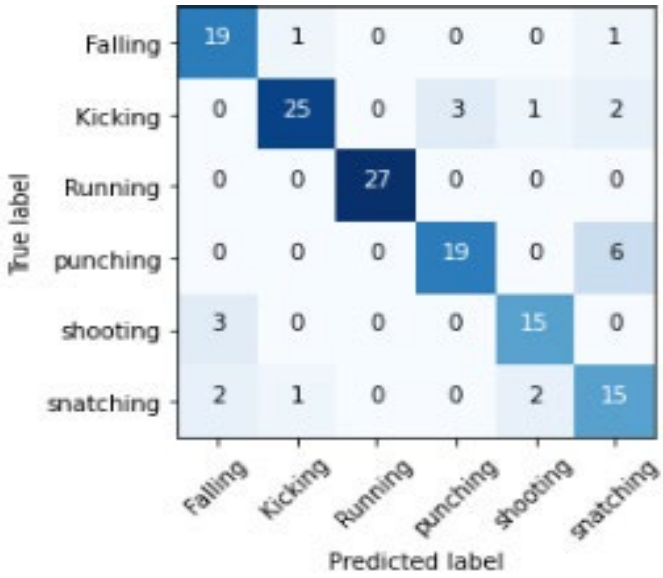


(a) Hybrid model accuracy



(b) Hybrid model loss

© IEOM Society International



(C) Hybrid Model Confusion Matrix

Figure 4. Hybrid Model Results Visualization.

11.1.2 Time Distributed Cnn Model

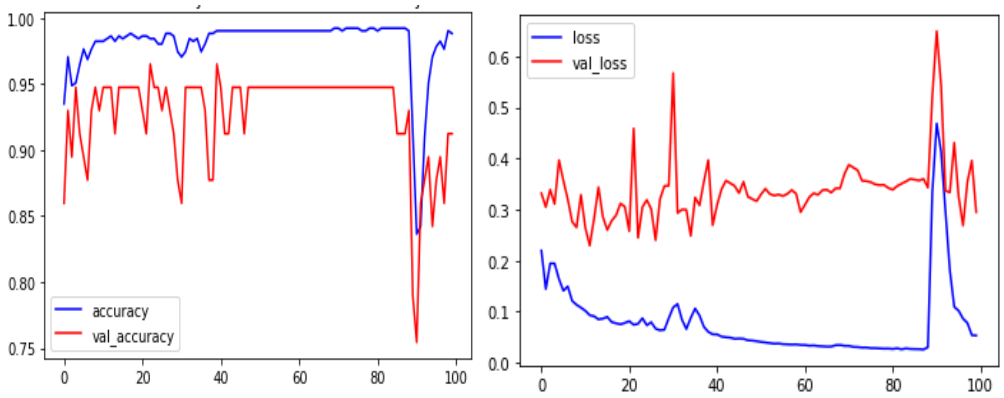
In contrast, the Time Distributed CNN model demonstrated outstanding results, achieving an accuracy of 90.14%. The precision, recall, and F1-score values were balanced, recorded at 90.78%, 90.14%, and 90.14%, respectively. This model utilizes a CNN architecture integrated with a time-distributed layer, enabling it to process each video frame individually while effectively capturing temporal dependencies across sequences. The study presents the complete training and evaluation process of the Time Distributed CNN model across multiple epochs, resulting in strong performance across all key metrics—accuracy, precision, recall, and F1-score. The model was trained for 100 epochs, during which continuous improvements were observed in its overall performance. The experimental results are illustrated through three visual representations: accuracy and loss curves (Figures 5a and 5b) and the confusion matrix (Figure 5c). These visual analyses reveal the model's increasing capability to learn and generalize effectively, as seen from the steady rise in accuracy and the reduction in loss over time. Additionally, the confusion matrix provides further insights into the model's classification proficiency across different categories, pinpointing any potential misclassification patterns. Overall, the findings emphasize the effectiveness of the Time Distributed CNN model in learning from training data and delivering highly reliable classification performance.

11.1.3 Keras Gru Model

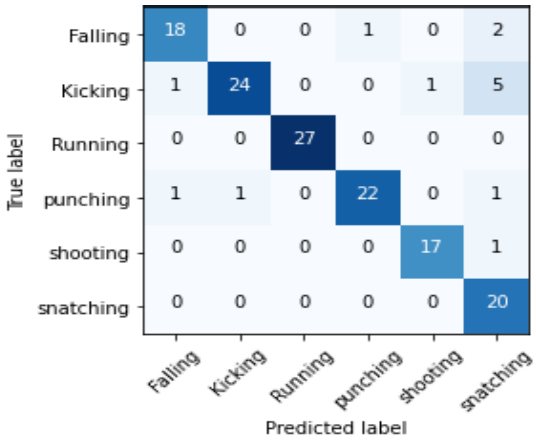
The Keras_GRU model achieved an accuracy of 83.80%, with corresponding precision, recall, and F1-score values of 87.10%, 84.10%, and 84.10%, respectively. This model employs Gated Recurrent Units (GRU) to effectively handle sequential data and exhibits performance that is competitive with the other tested architectures. The training progress of the Keras_GRU model is documented through detailed logs, reflecting accuracy and loss trends across multiple epochs for both training and validation phases.

The confusion matrix evaluates the model's classification accuracy across six activity categories: *kicking*, *falling*, *shooting*, *punching*, *running*, and *snatching*. In this matrix, each row represents the actual class, while each column indicates the predicted class. The model performs exceptionally well in identifying *running* activities, achieving nearly perfect classification for that category. However, it encounters challenges in accurately distinguishing instances of *snatching*, often misclassifying them as *kicking*, *falling*, or *shooting*. Some confusion is also observed between pairs such as *kicking* and *snatching*, and *falling* and *shooting*. A closer examination of the confusion matrix provides valuable insights into the model's areas of strength and limitation. Future improvements should focus on enhancing the model's discriminative ability to better differentiate between visually or temporally similar actions, particularly among the mentioned overlapping categories (Figure 5).

© IEOM Society International



(a) Distributed CNN Model Accuracy (b) Distributed CNN Model Loss



(b) Distributed CNN Model Confusion Matrix

Figure 5. Time Distributed Cnn Model Results Visualization.

11.1.4 Conv3d Model

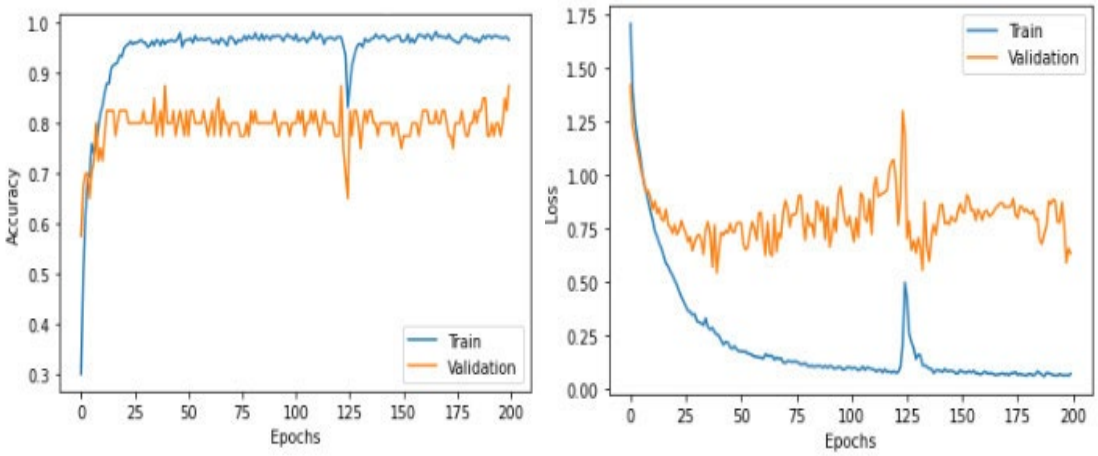
The Conv3D model achieved an accuracy of 88.23%, with precision, recall, and F1-score values all maintaining a consistent level of 88.20%. This model utilizes three-dimensional convolutional layers to directly process spatio-temporal information, demonstrating strong effectiveness in video-based activity classification tasks. As illustrated in Figure 7, the model was trained over 20 epochs, with each epoch representing one complete pass through the training dataset. Throughout training, the model's performance steadily improved, reflected by the rise in accuracy and the corresponding decline in loss for both training and validation data.

At the initial epoch, the training accuracy was 14.69% with a loss of 1.8210, while the validation accuracy measured 18.58% with a loss of 1.7199. By the 20th epoch, the model reached significantly higher accuracy levels—99.49% on the training data and 89.38% on the validation data—alongside substantially reduced loss values. This indicates that the model effectively learned to classify the video inputs with high precision.

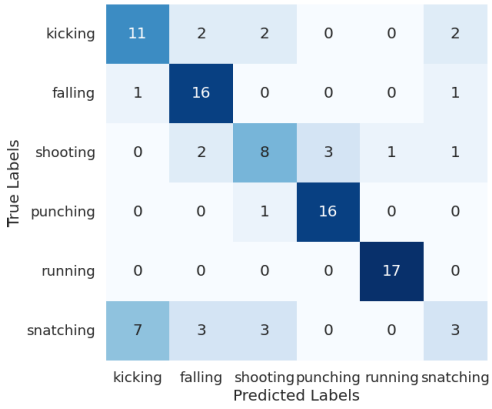
The confusion matrix further provides critical insights into the model's classification capabilities. It displays how frequently each true activity class (rows) is predicted as another class (columns). For instance, the model successfully recognizes most instances of *kicking* and *running*, but faces challenges with *snatching*, occasionally misclassifying it as *falling*, *punching*, or *shooting*. Such analysis helps in understanding the strengths and weaknesses of the Conv3D architecture in distinguishing between specific actions.

© IEOM Society International

The visual outputs-Conv3D Model Accuracy and Conv3D Model Confusion Matrix-complement the textual explanation by providing graphical representations of the model's learning progress and classification behavior over multiple epochs (Figure 6).



(a) Keras_GRU Model Accuracy (b) Keras_GRU Model Loss



(c) Keras_GRU Model Confusion Matrix

Figure 6. Keras_GRU Model Results Visualization.

In conclusion, the experimental outcomes reveal that each deep learning model varies in its efficiency for detecting and categorizing suspicious human activities in video data. Among all evaluated models, the Time Distributed CNN demonstrates the highest overall accuracy and most consistent performance.

12. Model Prediction

Model prediction involves leveraging a well-trained deep learning model to generate accurate outcomes or classifications for new and unseen data. In this context, the first phase focuses on making predictions using the trained model on test data that the model has not encountered during training. This test data, distinct from the training dataset, serves to evaluate the model's capability to generalize to unfamiliar scenarios.

As illustrated in Figure 8a, the model produces prediction outputs for a test video titled "newfi32 - 6.avi." It provides probability-based predictions for several possible actions occurring in the video, such as kicking, snatching, firing, punching, falling, and running, each accompanied by confidence scores. For instance, the model correctly identifies the action as "kicking" with a high confidence level of 99.12%, while assigning considerably lower probabilities to the other potential actions.

© IEOM Society International

In the subsequent phase, depicted in Figure 8b, the trained model is applied to a YouTube video to further demonstrate its predictive accuracy. In this case, the model successfully identifies the action "kicking" with a confidence score of 0.64% at the exact timestamp of 00:05. This prediction results from the model's analysis of video frames and classification based on learned spatial-temporal patterns and extracted features from the training phase. The associated confidence score indicates the model's level of certainty regarding its prediction.

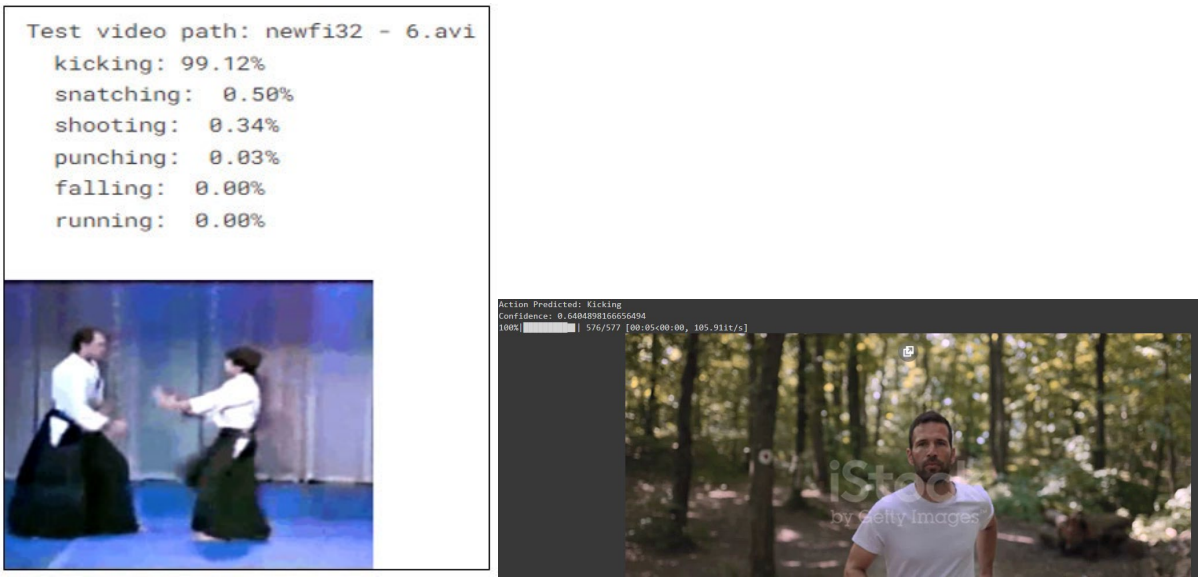
Overall, model prediction utilizes a trained neural network to interpret new data and produce reliable classifications or decisions founded on previously learned representations. This automated process proves particularly valuable in domains such as video surveillance, object detection, and natural language processing, where rapid and accurate interpretation of complex data is essential.

13. Comparative Analysis

The comparison presented in Table provides a comprehensive evaluation of both existing and proposed research approaches in this domain. The state-of-the-art methodology, as reported by Khan et al. (citekhan2022human), involves several models, including MLP (Multi-Layer Perceptron), CNN, LSTM, BiLSTM (Bidirectional LSTM), and CNN-LSTM, with the accuracy of each model reported as a percentage. In contrast, the proposed approach introduces novel models such as the Hybrid model, Time distributedTime, Keras_GRU, and Conv3D, each with its respective accuracy score.

When comparing the current and proposed methodologies, it is clear that the newly introduced models generally achieve higher accuracy. For example, while the existing CNN-LSTM model reaches a maximum accuracy of 76.50%, the proposed Time distributed Time model achieves a significantly higher accuracy of 90.14%. Similarly, the Conv3D model in the proposed approach shows substantial improvement over current techniques, attaining an accuracy of 88.23%.

This analysis highlights the effectiveness of the proposed models in delivering superior performance within the specific domain under study, compared to existing approaches. Such comparative evaluations are essential for guiding researchers and practitioners in selecting the most effective methods for their particular tasks or applications (Figure 7).



(a) Unseen test video prediction (b) Unseen YouTube video prediction

Figure 7. Model Prediction.

© IEOM Society International

Table 2. Comparative analysis of proposed and existing research.

	Model	Accuracy (%)
Existing Approach [25]	MLP	71.51
	CNN	75.47
	LSTM	66.09
	BiLSTM	66.26
	CNN-LSTM	76.50
Proposed Approach	Hybrid Model	84.51
	Time Distributed CNN	90.14
	Keras_GRU	83.80
	Conv3D	88.23

14. Conclusion and Future Work

This study presents a systematic approach for detecting suspicious human activity through a series of essential procedural steps. The research began with the collection of data from multiple sources, labeled as S1 and S2. This data was then carefully processed by eliminating inconsistencies, standardizing formats, and integrating it into a unified dataset. Comprehensive image preprocessing—including normalization, scaling, and augmentation—was performed to ensure uniformity and enhance image quality. Additionally, dataset annotation was carried out to improve the accuracy of classifying and predicting suspicious human behavior using supervised learning techniques. For model development and evaluation, the dataset was divided into training and testing sets to ensure unbiased assessment. CNNs were employed to extract key features from the preprocessed images, providing a foundation for subsequent model implementation. To detect suspicious human activity, multiple deep-learning architectures were explored, including the Hybrid LSTM model, time-distributed CNN, Keras_GRU, and Conv3D. Each model leveraged the features extracted by CNNs to identify patterns and generate predictions.

Our results demonstrated that the proposed time-distributed CNN model achieved a notably high accuracy of 90.14%, highlighting its effectiveness in detecting suspicious human activities. Similarly, the Conv3D model exhibited significant improvement over existing methods, attaining an accuracy of 88.23%. Finally, the trained models were applied to predict suspicious human actions in real-world scenarios, including analyzing YouTube videos and unexpected video footage. This practical validation emphasizes the relevance and utility of the proposed methodology in strengthening surveillance and security systems by improving the identification and mitigation of potential threats. Looking ahead, several directions exist for future research and enhancement. Expanding the proposed approach to include a wider variety of datasets, covering more types of suspicious activities and environmental contexts, could further improve model performance and adaptability. Exploring advanced deep-learning architectures, such as attention mechanisms and transformer-based models, may significantly enhance the accuracy and efficiency of activity recognition. Integrating real-time data streaming and processing capabilities into surveillance systems would enable faster detection and response to suspicious activities, thereby improving overall security operations.

Collaboration with domain experts and stakeholders is essential to increase the robustness and reliability of the proposed method. Such collaboration can aid in refining data annotation procedures and validating model predictions in real-world scenarios. Moreover, considering ethical concerns and privacy implications related to surveillance systems, future research should prioritize the creation of transparent and responsible frameworks for data collection, storage, and usage. This includes implementing strong data protection measures and complying with relevant regulations to safeguard individual privacy while maintaining effective security protocols. Addressing these challenges will help develop surveillance systems that are both ethically responsible and operationally efficient.

References

Rezaee, K., Rezakhani, S. M., Khosravi, M. R. and Moghimi, M. K., A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance, Personal and Ubiquitous Computing, vol. 28, no. 1, pp. 135–151, Feb. 2024.

Perez, M., Kot, A. C. and Rocha, A., Detection of real-world fights in surveillance videos, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2662–2666, May 2019.

Amrutha, C. V., Jyotsna, C. and Amudha, J., Deep learning approach for suspicious activity detection from surveillance video, Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 335–339, Mar. 2020.

© IEOM Society International

Sultani, W., Chen, C. and Shah, M., Real-world anomaly detection in surveillance videos, Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479–6488, Jun. 2018.

Wei, J., Zhao, J., Zhao, Y. and Zhao, Z., Unsupervised anomaly detection for traffic surveillance based on background modeling, Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 129–136, Jun. 2018.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Hassanien, A. E. and Pandey, H. M., An optimized dense convolutional neural network model for disease recognition and classification in corn leaf, Computers and Electronics in Agriculture, vol. 175, Art. no. 105456, Aug. 2020.

Teja, R., Nayar, R. and Indu, S., Object tracking and suspicious activity identification during occlusion, International Journal of Computer Applications, vol. 179, no. 11, pp. 29–34, Jan. 2018.

Ma, S., Sigal, L. and Sclaroff, S., Learning activity progression in LSTMs for activity detection and early detection, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1942–1950, Jun. 2016.

Varol, G., Laptev, I. and Schmid, C., Long-term temporal convolutions for action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

Ghazal, S., Khan, U. S., Saleem, M. M., Rashid, N. and Iqbal, J., Human activity recognition using 2D skeleton data and supervised machine learning, IET Image Processing, vol. 13, no. 13, pp. 2572–2578, 2019.

Zhu, G., Zhang, L., Shen, P. and Song, J., An online continuous human action recognition algorithm based on the Kinect sensor, Sensors, vol. 16, no. 2, p. 161, Jan. 2016.

Manzi, A., Dario, P. and Cavallo, F., A human activity recognition system based on dynamic clustering of skeleton data, Sensors, vol. 17, no. 5, p. 1100, May 2017.

Hbali, Y., Hbali, S., Ballihi, L. and Sadgal, M., Skeleton-based human activity recognition for elderly monitoring systems, IET Computer Vision, vol. 12, no. 1, pp. 16–26, Feb. 2018.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., Large-scale video classification with convolutional neural networks, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732, Jun. 2014.

Feichtenhofer, C., Pinz, A. and Zisserman, A., Convolutional two-stream network fusion for video action recognition, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1933–1941, Jun. 2016.

Li, J., Wu, R., Zhao, J. and Ma, Y., Convolutional neural networks (CNN) for indoor human activity recognition using ubisense system, Proceedings of 29th Chinese Control and Decision Conference (CCDC), pp. 2068–2072, May 2017.

Anishchenko, L., Machine learning in video surveillance for fall detection, Proceedings of Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), pp. 99–102, May 2018.

Gul, M. A., Yousaf, M. H., Nawaz, S., Ur Rehman, Z. and Kim, H., Patient monitoring by abnormal human activity recognition based on CNN architecture, Electronics, vol. 9, no. 12, p. 1993, Nov. 2020.

Ullah, W., Ullah, A., Haq, I. U., Muhammad, K., Sajjad, M. and Baik, S. W., CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks, Multimedia Tools and Applications, vol. 80, no. 11, pp. 16979–16995, May 2021.

Butt, U. M., Letchmunan, S., Hafinaz, F., Zia, S. and Baqir, A., Detecting video surveillance using VGG19 convolutional neural networks, International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, 2020.

Arshad, Q.-U.-A., Raza, M., Khan, W. Z., Siddiqa, A., Muiz, A., Khan, M. A., Tariq, U., Kim, T. and Cha, J.-H., Anomalous situations recognition in surveillance images using deep learning, Computers, Materials & Continua, vol. 76, no. 1, pp. 1103–1125, 2023.

Vrskova, R., Hudec, R., Kamencay, P. and Sykora, P., A new approach for abnormal human activities recognition based on ConvLSTM architecture, Sensors, vol. 22, no. 8, p. 2946, Apr. 2022.

Gandapur, M. Q. and Verdú, E., ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 4, p. 88, 2023.

Rajeswari, R., Anomalous human activity recognition from video sequences using BRISK features and convolutional neural networks, Galaxy International Interdisciplinary Research Journal, vol. 10, no. 2, pp. 269–280, 2022.

Khan, I. U., Afzal, S. and Lee, J. W., Human activity recognition via hybrid deep learning based model, Sensors, vol. 22, no. 1, p. 323, Jan. 2022.

© IEOM Society International