IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2024, accepted 29 July 2024, date of publication 1 August 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436653

RESEARCH ARTICLE

Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning

MONJI MOHAMED ZAIDI[1,2], GABRIEL AVELINO SAMPEDRO[3], (Member, IEEE), AHMAD ALMADHOR[4], SHTWAI ALSUBAI[5], ABDULLAH AL HEJAILI[6], MICHAL GREGUS[7], AND SIDRA ABBAS[8], (Graduate Student Member, IEEE)

[1]Department of Electrical Engineering, College of Engineering, King Khalid University, Abha 61421, Saudi Arabia
[2]Center for Engineering and Technology Innovations, King Khalid University, Abha 61421, Saudi Arabia
[3]School of Management and Information Technology, De La Salle—College of Saint Benilde, Manila 1004, Philippines
[4]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[5]College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia
[6]Faculty of Computers & Information Technology, Computer Science Department, University of Tabuk, Tabuk 71491, Saudi Arabia
[7]Faculty of Management, Comenius University in Bratislava, 82005 Bratislava, Slovakia
[8]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan

Corresponding authors: Michal Gregus (michal.gregus3@uniba.sk) and Sidra Abbas (sidraabbas@ieee.org)

This work was supported in part by the Deanship of Research and Graduate Studies, King Khalid University, through the Large Research Project RGP2/384/45; and in part by the Slovak Research and Development Agency under Contract APVV 19-0581.

ABSTRACT Suspicious Human activity recognition (SHAR) is crucial for improving surveillance and security systems by recognizing and reducing possible hazards in different situations. This study focuses on the task of precisely identifying potentially suspicious human behaviour by utilizing an innovative approach that harnesses advanced deep learning methods. Despite the abundance of research on the subject of SHAR, current methods frequently need to be revised with restricted levels of precision and efficiency. This research aims to address these constraints by presenting a thorough methodology for detecting and recognizing suspicious human activities. By rigorously collecting and preparing data, as well as training models, we aim to solve the issue of inaccurate and inefficient activity recognition in surveillance systems. By utilizing Convolutional Neural Networks (CNNs) and deep learning structures, such as the proposed time-distributed CNN model and Conv3D model, we attain notably enhanced accuracy rates of 90.14% and 88.23%, respectively, surpassing current research approaches. Moreover, the efficacy of our approach is illustrated by conducting prediction experiments on previously unreported test data and YouTube videos. Through the process of evaluating the trained models on unseen test data, we ascertain their accuracy and ability to apply learned knowledge to new situations. Moreover, the algorithms are utilized to predict dubious human conduct in a YouTube video, demonstrating their practical usefulness in real-life surveillance situations. The results of this study have important consequences for improving surveillance and security systems, allowing for better identification and reduction of possible dangers in various settings. Our methodology enhances the precision and effectiveness of SHAR, advancing the construction of more resilient and dependable surveillance systems and ultimately strengthening public safety and security.

INDEX TERMS Suspicious human activity recognition (SHAR), deep learning, convolutional neural network, multimedia data.

I. INTRODUCTION

The widespread incorporation of many applications in modern society has significantly transformed many aspects of our lives, with visual systems emerging as essential instruments. One important area of study in this field is the detection of suspicious human behaviour using video surveillance, which involves classifying behaviours as either normal or abnormal [1]. The increasing frequency of disruptive incidents in public areas globally, ranging from

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

banks to airports, highlights the urgent requirement for efficient security measures [2]. As a result, surveillance systems, mostly dependent on CCTV cameras, have grown quite common, producing large quantities of video data for examination. Nevertheless, the labour-intensive nature of manual monitoring makes it unfeasible, thus necessitating the development of automated detection systems [3].

Researchers are using breakthroughs in machine learning, artificial intelligence, and deep learning to improve surveillance systems. Their goal is to proactively identify and categorize suspicious activity [4]. The objective of this project is to implement deep learning models for the purpose of identifying and categorizing six primary activities: Running, Punching, Falling, Snatching, Kicking, and Shooting. This will enhance security measures and allow for prompt intervention [5]. Deep learning architectures, specifically CNNs, have emerged as strong tools for extracting essential capabilities from video data aimed toward facilitating efficient detection [6]. Yakkali et al. [7] suggested the utilization of digital image and video processing techniques to monitor item movement. They underscore the importance of training deep temporal models for accurate activity identification, as emphasized by Ma et al. [8]. Their emphasis lies in highlighting the importance of Recurrent Neural Networks (RNNs), mainly long short-term memory (LSTM) models, in comprehending the progression of activities and minimizing classification errors. Moreover, improvements in video representation learning, in particular in long-term Temporal Convolutions (LTC), demonstrate promise in improving activity recognition [9]. However, there persists a need to enlarge the scope of detectable activities and improve overall performance metrics.

A. RESEARCH CONTRIBUTION

- Presented a novel approach for detecting potentially suspicious human behaviour by leveraging deep learning techniques. The efficacy of the suggested methodology is validated by conducting prediction scenarios using previously unseen test data and by creating a YouTube video.
- To address the limited availability of publicly accessible data on this subject, a new dataset has been created. This dataset consists of surveillance videos from various sources and includes 6 different physical activities. A thorough comparative analysis is conducted to determine the most effective model for activity recognition.
- By utilizing Convolutional Neural Networks (CNNs) and sophisticated deep learning structures, such as the suggested Time Distributed CNN model and Conv3D model, we attain notably enhanced accuracy rates of 90.14% and 88.23%, respectively, surpassing current research approaches.

The following sections of this paper are organized to provide a more detailed examination of the background and relevant literature (Section II), clarify the research methodology (Section III), provide the dataset details (Section IV), and analyze the findings (Section V). Ultimately concluding with a summary and future work (Section VI). This study seeks to make a valuable contribution to the continuing discussion on video surveillance and activity detection. The findings of this study have the potential to strengthen security measures and improve public safety.

II. LITERATURE REVIEW

While Human activity recognition has been a topic of considerable study in present literature, this section delves into the latest improvements in this domain. Cutting-edge studies in Human activity recognition predominantly revolve across the realms of machine learning and deep learning methodologies.

In the area of machine learning, Ghazal et al. [10] carried out a comparative study specializing in Human activity recognition with the use of 2d-skeletal facts. They utilized the OpenPose package to extract visual and motion attributes from 2d landmarks of human skeletal joints. The examine evaluated five supervised machine learning strategies, consisting of support Vector machine (SVM), Naive Bayes (NB), Linear Discriminant (LD), k-nearest neighbours (KNNs), and feed-forward backpropagation neural networks. The primary objective was to identify four awesome activity instructions: sitting, standing, walking, and falling, with the k-nearest neighbours (KNNs) exhibiting the most promising overall performance. In another study, Zhu et al. [11] introduced an online continuous Human action recognition (CHAR) approach based on skeletal records captured from Kinect depth sensors. Their approach employed a variable-length maximum Entropy Markov model (MEMM) for continuous hobby reputation without the need for earlier detection of activity begin and give-up factors. Additionally, a singular technique utilized bone information from a depth camera, leveraging machine learning to identify human actions appropriately [12].

In comparison to previous methodologies where every activity is identified by using a unique range of clusters distinct from activity instances, Hbali et al. [13] proposed a unified method based on skeletons to analyze the spatial-temporal elements of human activity sequences. Their approach concerned using Minkowski and cosine distances to quantify the dissimilarity between joint data acquired from Microsoft Kinect. The model was trained and assessed using publicly available datasets such as MSR each day activity 3D and Microsoft MSR 3D motion. Leveraging the extremely Randomized Tree technique, the model showcased promising results in the improvement of monitoring systems for the elderly. Considerably, this was achieved through the utilization of low-cost depth sensors and open-source libraries.

Karpathy et al. [14] underscored the effectiveness of CNNs in addressing challenges related to image identification. Their study focused on the classification of a diverse array of videos, leveraging a dataset comprising 1 million videos categorized into 487 various categories. Exploring numerous strategies to comprise local spatio-temporal information into CNNs, they proposed a multi-resolution architecture aimed at expediting the training process. Through the retraining of the top layers of the model using the UCF-one hundred and one action recognition dataset, researchers determined enormous enhancements within the model's generalization competencies, resulting in an incredible growth in accuracy from the baseline model's 43.9% to 63.3%. Feichtenhofer et al. [15] investigated the current improvements in utilizing CNNs for detecting human activities in videos. Their focus changed to strategies that remember both the visible appearance and the actions of subjects. The study delved into leveraging CNN towers to harness spatio-temporal facts, highlighting the potential of merging spatial and temporal networks at a convolution layer without compromising performance. Introducing a modern CNN architecture for integrating video capabilities throughout both space and Time, the researchers established exceptional overall performance on broadly recognized evaluation datasets.

Li et al. [16] proposed a recognition method using a CNN to enhance the accuracy of indoor human activity identification using geographical location data. Their state-of-the-art system, comprising convolutional layers, fully connected layers, and max pooling, performed high-quality consequences via appropriately identifying six behaviours with a recognition rate of 86.7%, demonstrating the practicality of their method. Anishchenko [17] investigated the software of deep studying and switch learning techniques for fall detection by studying information captured from security cameras. Utilizing the CNN AlexNet architecture, the classifier becomes tailor-made to detect falls especially. The purpose was to enhance its efficiency by incorporating new heuristics that consider the temporal association of frames and the typical period of fall activities. Gul et al. [18] delved into the utilization of the You Look Only Once (YOLO) network as the primary CNN model for real-time affected person surveillance geared toward spotting human actions. Through retraining the model across 32 epochs using categorized affected person behaviour snapshots, researchers achieved an impressive accuracy of 96.8% in action recognition, underscoring the capacity of their technique. In their research, Ullah et al. [19] brought an advanced anomaly detection framework leveraging a pre-trained CNN model for function extraction from video frames observed via processing with BD-LSTM. Their architecture verified promising results on the UCF-Crime dataset, illustrating the functionality for effective anomaly identity in surveillance networks.

Our proposed work aims to identify six suspicious acts (Running, Punching, Falling, Snatching, Kicking, and Shooting) that have not been previously examined in the existing literature. An automatic video detection system is essential because of the difficulties associated with continuously monitoring camera images in public areas. Although earlier studies have achieved significant progress in the field of intelligent surveillance, the task of reaching flawless detection and accuracy rates continues to be a challenging endeavour [20].

III. PROPOSED APPROACH

The proposed methodology for identifying suspicious human activity entails several crucial stages, as shown in Figure 1. Initially, data information is gathered from distinct sources, denoted as S1 and S2. This data then undergoes meticulous preparation regarding cleansing, formatting, and integration to produce a cohesive dataset that consolidates relevant information from both sources. Significant interest is given to the preparation of images within the dataset. Techniques consisting of normalization, scaling, and augmentation are employed to ensure uniformity and enhance the excellent image inputs for the next evaluation. Furthermore, the dataset is annotated, with instances of suspicious human behaviour being categorized to facilitate accurate classification and prediction by using supervised learning algorithms. To facilitate the training and assessment of the model, the dataset is partitioned into separate units for training and testing functions. This ensures that the model is skilled in an extensive extent of data while retaining a distinct subset of data for rigorous testing and evaluation. This division guarantees the integrity of the evaluation process by evaluating the performance of the model on data that has not been previously viewed. CNNs are then utilized to extract significant characteristics from the preprocessed image data. These characteristics act as informative representations of the input data and are crucial in later model implementations.

Several deep learning architectures, such as the Hybrid LSTM Model, Time Distributed CNN, Keras_GRU, and Conv3D, are used to tackle the task of identifying suspicious human activity. Every model utilizes the extracted CNN-based features to acquire knowledge of patterns and provide predictions. Ultimately, the trained models are implemented to predict potentially dubious human behaviours in real-life situations, encompassing the anticipation of YouTube videos and the examination of unfamiliar video data. This step highlights the practical usefulness and ability to be applied in various situations involving the established strategy. In general, this methodology seeks to improve surveillance and security systems by increasing the ability to identify and stop possible threats.

IV. DATASET SELECTION

The dataset component of the research paper outlines the methodology for developing a dataset customized specifically for video classification. Due to the lack of available datasets that meet the specific needs of the study, the researchers proactively gathered and organized their datasets. The dataset consists of videos that depict six unique categories: Falling, Kicking, Running, Punching, Shooting, and Snatching. In order to compile this dataset, videos were

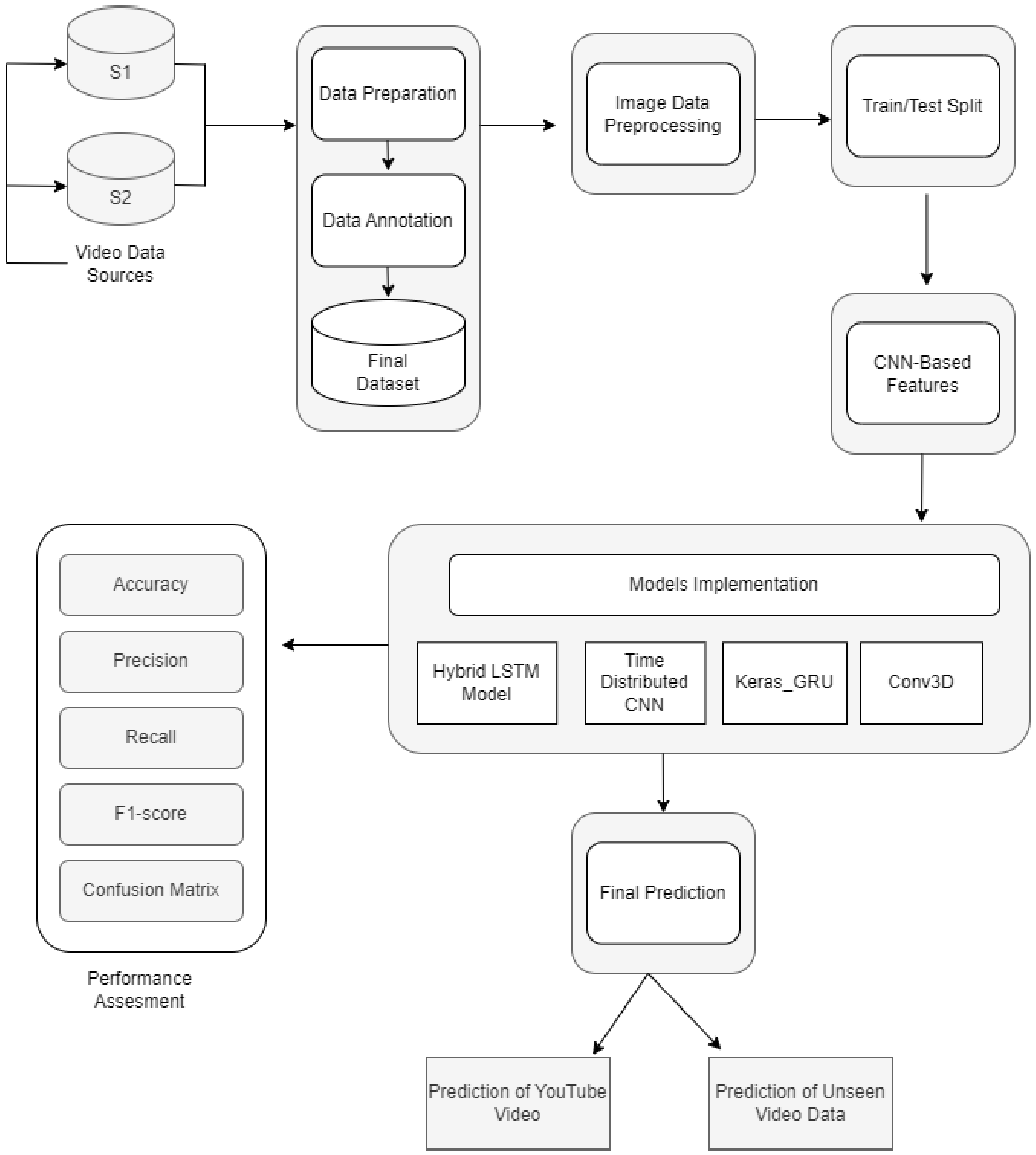


FIGURE 1. Proposed Approach for Suspicious Human Activity Recognition.

obtained from two web platforms, specifically Sources (S1,[1] S2[2]). The selection of these platforms was based on their ease of access and the presence of videos that are pertinent to the desired courses.

After obtaining the videos, they were sorted into separate files based on their assigned categories. As an illustration, movies that showed falling behaviour were organized in a folder labelled "falling," whereas recordings that displayed kicking activities were placed in a folder titled "kicking." This organization enabled the methodical administration and retrieval of videos during the following phases of data preparation and model training. The method of creating the dataset yielded a total of 564 videos specifically allocated for training and 142 videos specifically set aside for testing. The class distribution can be seen in Figure 2. The allocation of videos among the six classes exhibited minor variations, with each class comprising a distinct number of video samples. The "snatching" class contained 113 videos, but the "kicking" class contained 119 videos. The presence of

[1] https://www.istockphoto.com
[2] https://www.videvo.net

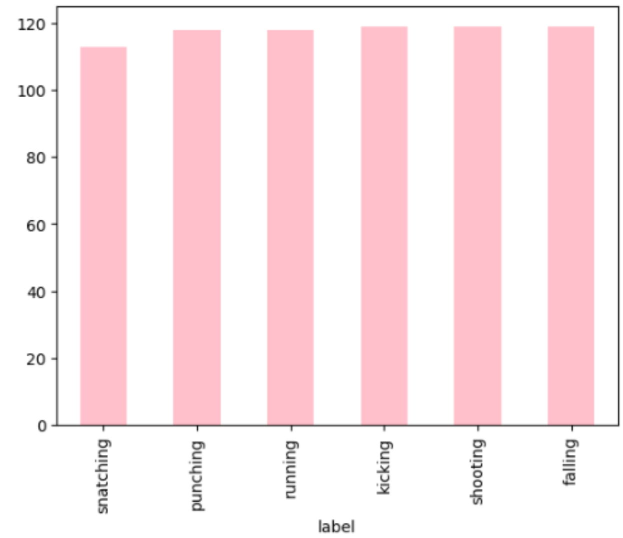


FIGURE 2. Dataset Distribution.

distinct class representations highlights the significance of maintaining a well-balanced dataset to mitigate biases and improve the model's capacity to generalize across diverse classes.

A. DATASET ANNOTATION

The process of data annotation was automated using Python code to enhance efficiency. The code sequentially scans all video files stored in the specified directory. Afterwards, it takes and stores the name of each video in a cache while also adding this information to a CSV file. This method simplified the process of data annotation by allowing the merging of all video file names into a single-label CSV file. After completing the data annotation, the dataset underwent additional processing by combining all relevant files to provide a consolidated label.csv file. This extensive file played a crucial role in the later stages of preparing the dataset and training the model. In addition, the dataset was divided into separate training and testing subsets to ensure that there were independent datasets available for both model creation and evaluation.

B. DATA PRE-PROCESSING

Data preprocessing serves as a foundational step in research implementation, especially in tasks related to video record evaluation. In this study, video information was preprocessed using the Python OpenCV library. The method is initiated by developing an empty list, acting as a repository for the extracted frames. Each video file is accessed through its corresponding caption, permitting retrieval of the total frame to be counted in the video. To systematically extract frames for evaluation, a specific interval was established based on a predetermined series length, set at 20 frames in this example. This interval calculation is derived by dividing the overall frame count by the sequence length, ensuring frames are extracted at ordinary intervals to cover the video content comprehensively. Upon extraction, every frame underwent resizing to standardize dimensions, both to 224 × 224 pixels or 64 × 64 pixels, based on specific analysis necessities. This resizing step ensured uniformity in frame size across all samples, facilitating constant processing and evaluation.

C. FEATURE EXTRACTION

For feature extraction, we harnessed the power of the InceptionV3 model, a custom variation of CNNs famous for its excellent performance in image analysis responsibilities. Our adaptation of the InceptionV3 model integrated specific parameters tailored to our requirements. These included leveraging pre-trained weights from the ImageNet dataset, excluding the top layer to facilitate feature extraction, implementing average pooling for dimensionality reduction, and adjusting the input shape to align with the dimensions of the video frames under consideration. Through this configuration, the InceptionV3 model systematically scrutinized each video frame, extracting salient features critical to our classification goals. Leveraging the hierarchical representations encoded within the InceptionV3 structure, we collected a complete set of 2048 awesome attributes for each frame. The structure of the InceptionV3 model for feature extraction is illustrated in Figure 3, showcasing its elaborate community structure devised to seize and encode problematic patterns and characteristics inherent in the video frames. This feature extraction method aimed to distil vital visible information from the video data, facilitating streamlined evaluation and classification in subsequent stages.

D. MODELS IMPLEMENTATION

Detecting suspicious human activities inside video recordings relies heavily on the robustness of deep learning architectures. Leveraging improvements in neural network structures, we explore a spectrum of models meticulously designed to capture the dynamic spatio-temporal patterns inherent in video sequences. This section entails a comprehensive examination of the implementation intricacies of high-quality deep learning architectures, together with the Hybrid LSTM version [21], Time distributTimeNN [22], Keras_GRU [23], and Conv3D [24]. Each model possesses unique strengths and abilities in activity popularity, using various techniques for characteristic extraction, series modelling, and type. Through rigorous experimentation and assessment, our objective is to identify the maximum optimal model architecture that excels in correctly discerning and categorizing suspicious movements within video record streams. The precise parameters of every model are outlined in Table 1.

E. PROPOSED WORK ALGORITHM

Algorithm 1 outlines the systematic approach employed in this study, providing a structured framework for conducting the research by delineating the specific steps undertaken throughout the investigative process. In the initial phase,

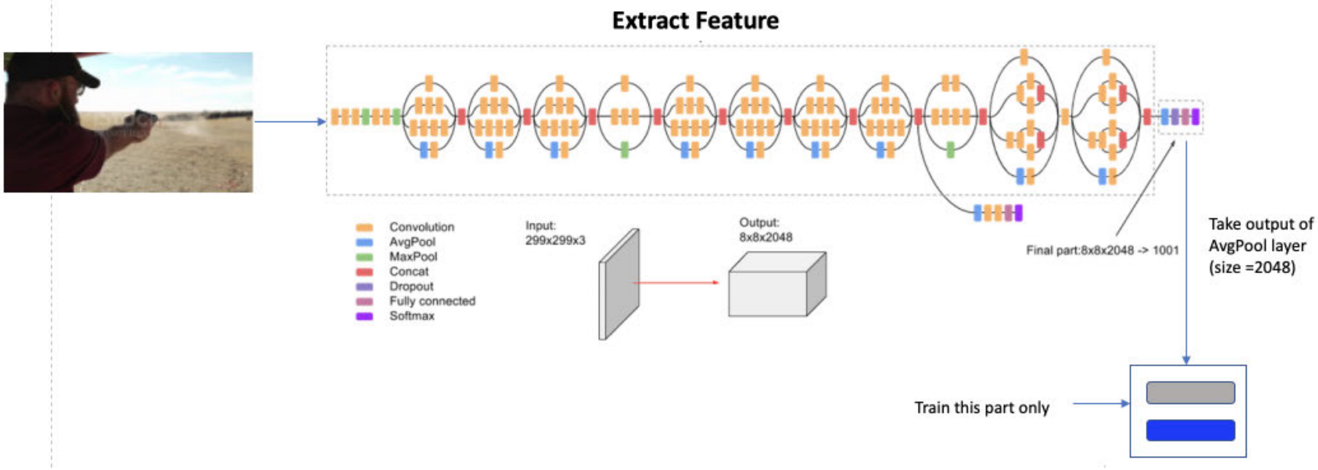


FIGURE 3. Features Extraction Method.

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

denoted as "Dataset Selection," data is sourced from two distinct origins, designated as S1 and S2. This preliminary stage involves the collection of relevant study data, establishing the foundational groundwork for subsequent analyses. Subsequently, in the phase termed "Data Preparation," the collected data undergoes cleaning, structuring, and integration procedures to construct a cohesive dataset. Through this process, data integrity is ensured, rendering it amenable to further processing and analysis. Following data preparation, the algorithm progresses to "Image Data Preprocessing," wherein image data undergoes preprocessing techniques such as scaling, augmentation, and normalization. These preprocessing steps enhance the quality and consistency of the image inputs, priming them for subsequent analysis.

The algorithm then proceeds to "Data Annotation," where instances of suspicious human behaviour within the dataset are annotated. This labelling facilitates the accurate classification and prediction of such events by supervised learning algorithms, streamlining the process of identifying and analyzing anomalous activities. The dataset is split into training and testing sets inside the "train/test split" stage, with 80% of the information unique for training and 20% for testing. The assessment of model performance on unobserved data is ensured via this segmentation. After that, the algorithm concentrates on "CNN-primarily based features," which are features derived from the preprocessed picture data using CNNs. For the purpose of implementing the model later on, these capabilities are characteristic of informative representations of the input data. The following step is performance assessment, when different measures are used to analyze the model's performance, including accuracy, precision, recall, F1-score, and confusion matrix. These measures shed light on the stability and efficacy of the models that have been put into practice.

The next step is "Models Implementation," wherein a number of deep learning models are put into practice, such as the Time Distributed CNN, Keras_GRU, Conv3D, and Hybrid LSTM Model. Every model is created to make the most of the features that have been extracted and to meet the goals of the research successfully. The process ends with "Final Prediction," in which unseen video data and suspicious activity in YouTube videos are predicted using the trained models. This step shows how the research findings can be used practically in real-world situations.

Algorithm 1 Proposed Algorithm for SHAR

```
1: Dataset Selection: Obtain data from sources $S1$ and $S2$
2: Data Preparation:
3:     Clean, format, and integrate data to create the final dataset
4: Image Data Preprocessing:
5:     Normalize, scale, and augment image data
6: Data Annotation: Label suspicious human behaviour instances in the dataset
7: Train/Test Split:
       Training Set = 80% of the dataset
       Testing Set = 20% of the dataset
8: CNN-Based Features:
       $F_1 = \text{Conv1}(X)$
9: Performance Assessment: Evaluate models using metrics:
10:    Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
11:    Precision: $\frac{TP}{TP+FP}$
12:    Recall: $\frac{TP}{TP+FN}$
13:    F1-score: $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
14: Models Implementation: Implement various deep learning models:
15:    Hybrid LSTM Model: $h_t = \text{LSTM}(x_t, h_{t-1})$
16:    Time Distributed CNN: $y = \text{TD-CNN}(X)$
17:    Keras_GRU: $h_t = \text{Keras\_GRU}(x_t, h_{t-1})$
18:    Conv3D: $y = \text{Conv3D}(X)$
19: Final Prediction:
20:    Predict suspicious activity in YouTube video
21:    Predict suspicious activity in unseen video data
```

V. EXPERIMENTAL ANALYSIS AND RESULTS

in this phase, we delve into the experimental effects stemming from the deployment of diverse deep-learning models aimed at discerning and categorizing suspicious human behaviours depicted in video content material. The overarching objective revolves around analyzing frames extracted from videos to facilitate the identity and category of numerous activities into six awesome classes. The models harnessed for this challenge embody Time distributTimeNN, Keras_GRU, Hybrid model, and Conv3D, each meticulously examined and evaluated to gauge its efficacy and overall performance in accurately spotting and classifying a spectrum of dubious activities captured in video datasets.

A complete assessment of each model's performance is performed, employing various evaluation metrics consisting of accuracy, precision, recall, F1-rating, and a detailed exam of the confusion matrix. To ensure rigorous assessment, the dataset is to start partitioned into wonderful training and testing subsets. Mainly, 564 videos are earmarked for training functions, imparting enough corpus of information for model training, while a special set of 142 videos is reserved totally for rigorous checking out and evaluation. By following this division, our goal is to uphold the integrity of the experimental process and precisely evaluate the performance of the trained models on data that has not been previously observed.

A. EXPERIMENTAL CONFIGURATION

The analysis utilized Python 3.8 and the Kaggle IDE. Given the time-consuming process of training deep learning models, it is crucial to ensure that the library installation is done correctly in order to achieve successful model training and execution. TensorFlow is highly regarded as one of the most commonly utilized libraries for creating effective image-processing models. The necessary libraries installed for this experiment comprised TensorFlow, Keras, Scikit-learn, Matplotlib, Pillow, and OpenCV. Keras functions as a leading tool for creating deep learning models and functions as a library with a high-level interface for the TensorFlow framework. By utilizing Scikit-learn, a Python toolbox, one may effectively employ machine learning techniques such as classification and regression. The Matplotlib library in Python is quite useful for visualizing data. Both OpenCV and Pillow are utilized for image-processing jobs. It is crucial to install all of these libraries in order to experiment successfully successfully.

B. EXPERIMENTAL RESULTS

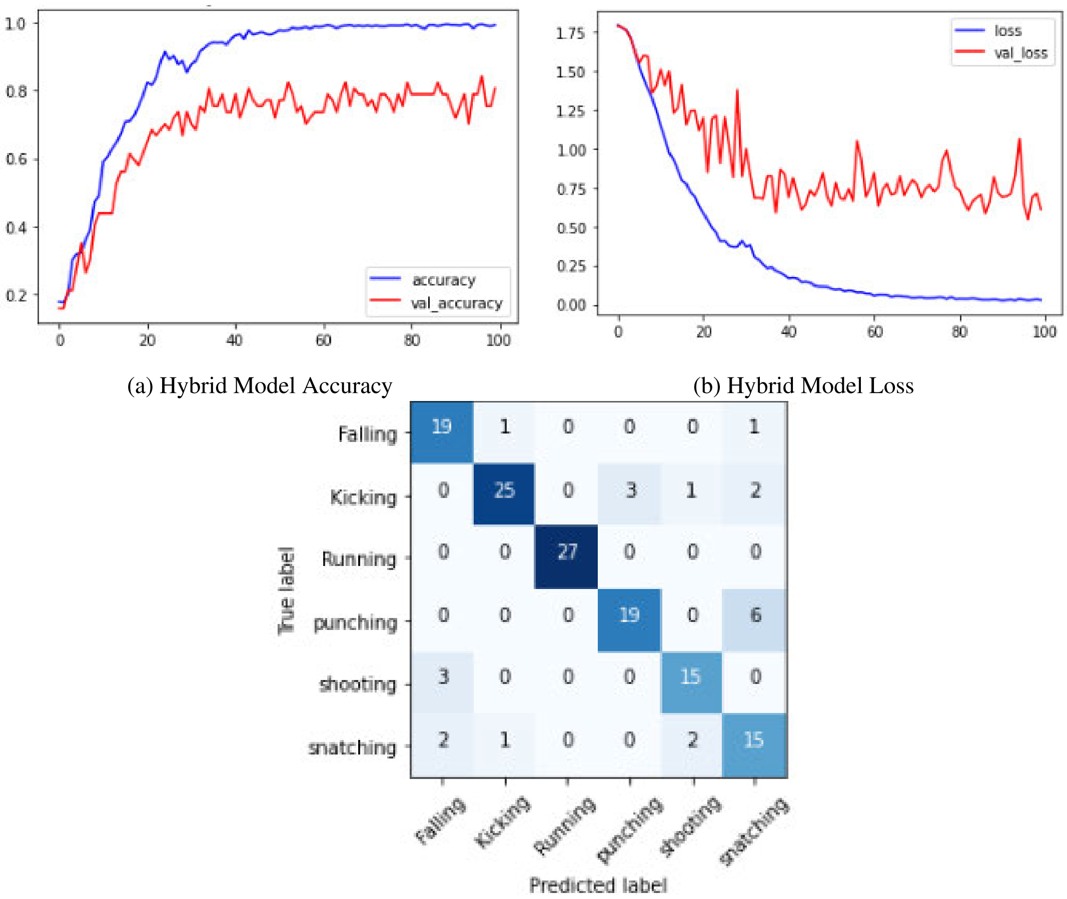
Table 2 displays the numerical outcomes derived from assessing several models on the dataset. The performance measures for each model, such as accuracy, precision, recall, and F1-score, are provided.

1) HYBRID MODEL

The Hybrid Model attained an accuracy of 84.51%, along with precision, recall, and F1-score values of 84.89%, 83.10%, and 84.72% correspondingly. This model integrates both CNN and LSTM architectures to exploit the spatial and temporal information present in the video input. During the training process of the hybrid model (Figure 4a), we notice a progressive rise in accuracy across epochs, suggesting enhanced performance as the model gains knowledge from the training data. Nevertheless, the validation accuracy seems to stabilize after approximately 25 epochs, indicating that the model may have hit its maximum performance level. Simultaneously, the loss (Figure 4b) consistently diminishes, suggesting that the model's predictions are improving in accuracy. The drop is observed in both the training and validation data, suggesting that the model is not suffering from overfitting. In addition, the confusion matrix (Figure 4c) offers valuable information about the model's performance in various classes, indicating regions where the model may face challenges in making precise predictions.

2) TIME DISTRIBUTED CNN MODEL

On the other hand, the Time Distributed CNN model exhibited exceptional performance, with an accuracy rate of 90.14%. The exhibited values for precision, recall, and



(a) Hybrid Model Accuracy

(b) Hybrid Model Loss

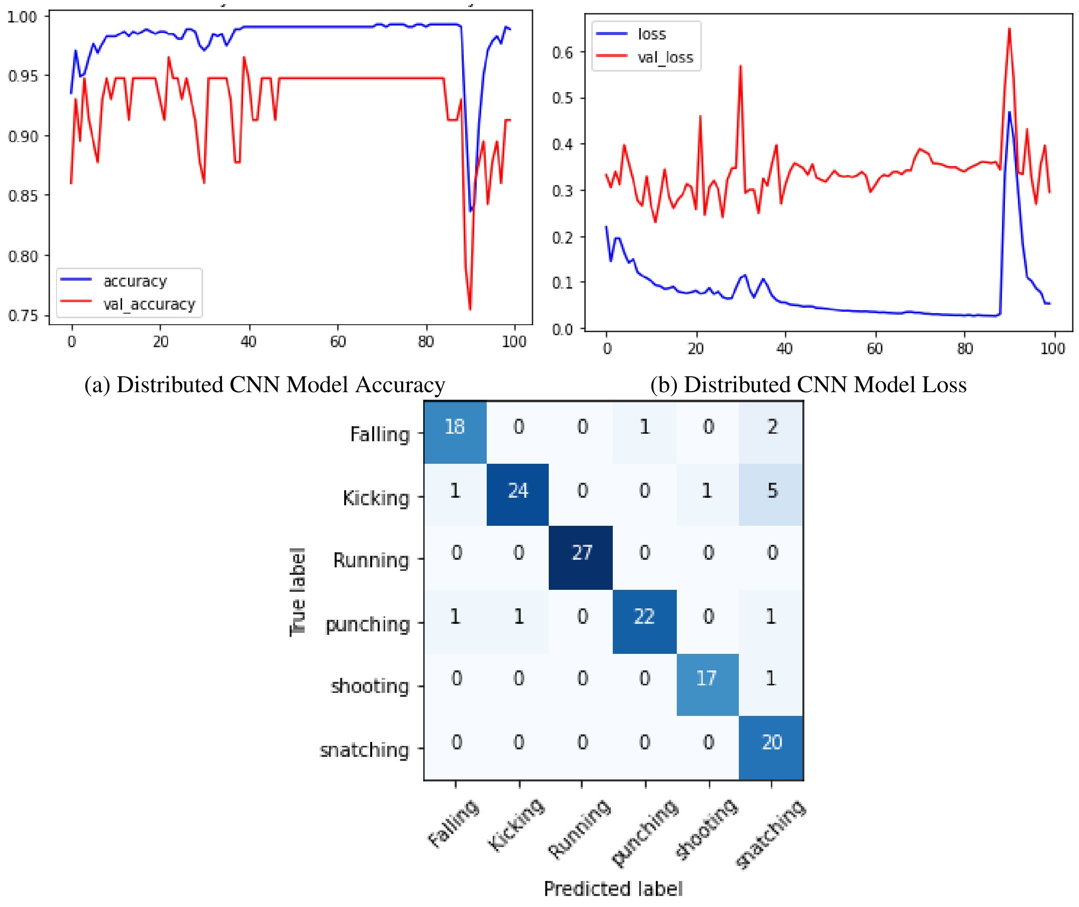
(c) Hybrid Model Confusion Matrix

FIGURE 4. **Hybrid Model Results Visualization.**

F1-score were all balanced and equal to 90.78%, 90.14%, and 90.14% correspondingly. This model employs a CNN architecture with a time-distributed wrapper, which allows it to handle each frame separately and effectively capture temporal information. This research outlines the training and testing of a Time disbursed CNN model throughout more than one epoch to classify information effectively, attaining excellent levels of accuracy, precision, recall, and F1 rating metrics. The training regimen spans 100 epochs, at some point of which there is a discernible development in overall performance metrics. The assessment effects are visually represented in 3 plots: accuracy and loss (Figure 5a and Figure 5b), at the side of a confusion matrix (Figure 5c). Through this training method, valuable insights emerge concerning the model's ability to examine and generalize successfully, obvious from the revolutionary enhancement in accuracy and reduction in loss over successive epochs. Moreover, the confusion matrix gives insights into the model's classification accuracy throughout specific classes, identifying any ability instances of misclassification. In summary, the consequences underscore the Time Distributed CNN model's efficacy in learning from the training statistics and demonstrating promising overall performance in categorizing the goal records.

3) KERAS GRU MODEL

The Keras_GRU model attained an accuracy of 83.80%, along with precision, recall, and F1-score values of 87.10%, 84.10%, and 84.10% correspondingly. This model utilizes GRU to process sequential data, demonstrating a competitive level of performance compared to other models. This research provided the training records of the Keras_GRU model. The logs display the training progress across numerous epochs, encompassing loss and accuracy data for both training and validation. These metrics are presented in Figures 6a and 6b. The confusion matrix depicted in Figure 6c presents an evaluation of the classification model's performance across six distinct categories: "kicking," "falling," "shooting," "punching," "running," and "snatching." The rows in the table correspond to the true class labels, while the columns reflect the anticipated class labels. The model demonstrates top-notch accuracy in predicting the activity of "running," effectively classifying all instances. However, it encounters demanding situations in accurately discerning



(a) Distributed CNN Model Accuracy

(b) Distributed CNN Model Loss

True label \ Predicted label	Falling	Kicking	Running	punching	shooting	snatching
Falling	18	0	0	1	0	2
Kicking	1	24	0	0	1	5
Running	0	0	27	0	0	0
punching	1	1	0	22	0	1
shooting	0	0	0	0	17	1
snatching	0	0	0	0	0	20

(c) Distributed CNN Model Confusion Matrix

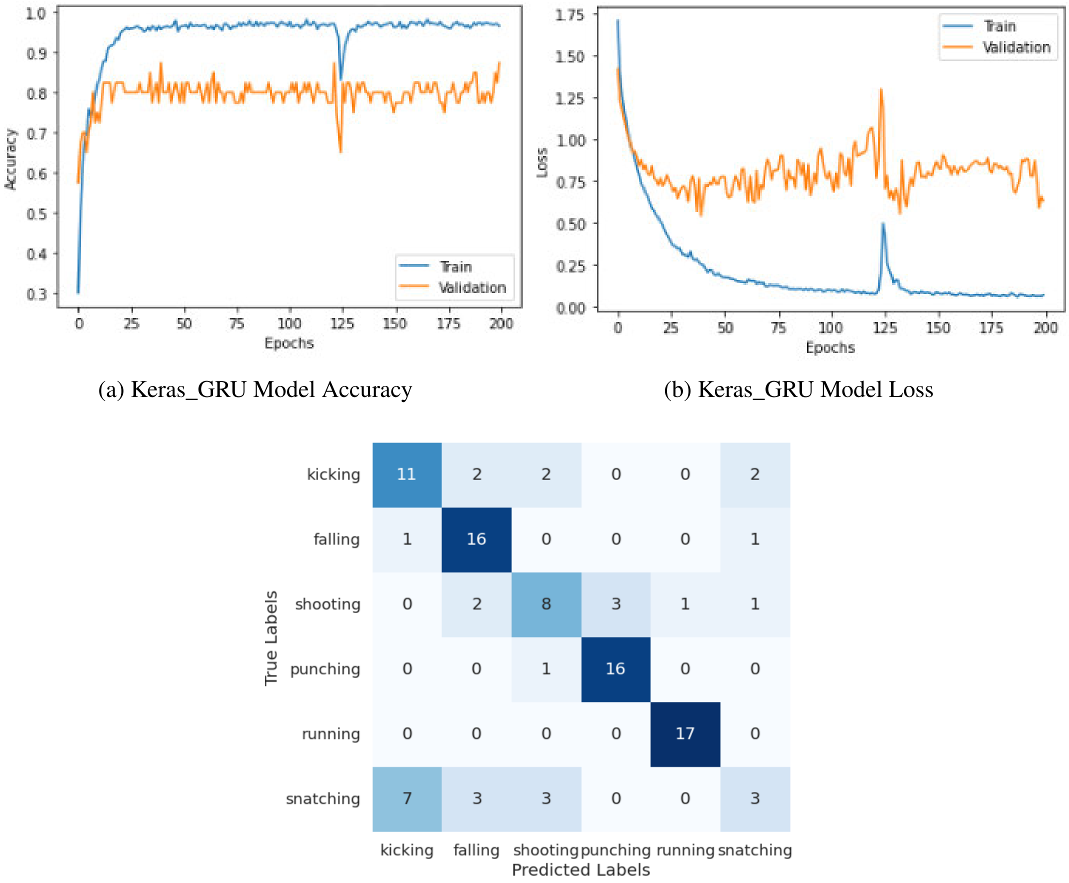
FIGURE 5. Time Distributed CNN Model Results Visualization.

times of "snatching," regularly misclassifying them as "kicking," "falling," or "shooting." additionally, there are times of ambiguity among categories such as "kicking" and "snatching," in addition to "falling" and "shooting." an in-depth exam of the confusion matrix offers precious insights into the model's strengths and weaknesses, guiding future efforts aimed at enhancing its performance. Particularly, efforts should focus on enhancing the model's ability as it should be differentiated among similar behaviours, along with "kicking" and "snatching," as well as "falling" and "shooting".

4) CONV3D MODEL

The Conv3D model achieved an accuracy of 88.23%, with consistent precision, recall, and F1-score values of 88.20%. This model employs three-dimensional convolutional layers to analyze spatio-temporal input directly, showcasing strong performance in video classification challenges. The output in Figure 7 demonstrates the training procedure of a Conv3D model for a total of 20 epochs. An epoch corresponds to a single iteration over the full training dataset. The performance of the model steadily increases across epochs, as seen by the increasing accuracy and decreasing loss values observed in both the training and validation sets. At epoch 1, the accuracy on the training set is 14.69% with a loss of 1.8210. On the validation set, the accuracy is 18.58% with a loss of 1.7199. Nevertheless, by epoch 20, the model attains notably superior accuracy on both the training (99.49%) and validation (89.38%) datasets, accompanied by substantially reduced losses. This indicates that the model has successfully acquired the ability to categorize the incoming data.

The confusion matrix offers valuable insights into the model's efficacy in categorizing various activities. The diagram depicts the frequency with which each real class (represented by rows) was predicted as each class (represented by columns). For example, it accurately categorizes the majority of occurrences of "kicking" and "running," but encounters difficulties when dealing with "snatching," incorrectly categorizing certain occurrences as "falling," "punching," and "shooting." This information is crucial for comprehending the model's capabilities and limitations in identifying particular actions. The given explanation references the images in the text using the following citations: "Conv3D Model Accuracy" (Figure 7a), "Conv3D



(a) Keras_GRU Model Accuracy

(b) Keras_GRU Model Loss

(c) Keras_GRU Model Confusion Matrix

FIGURE 6. **Keras_GRU Model Results Visualization.**

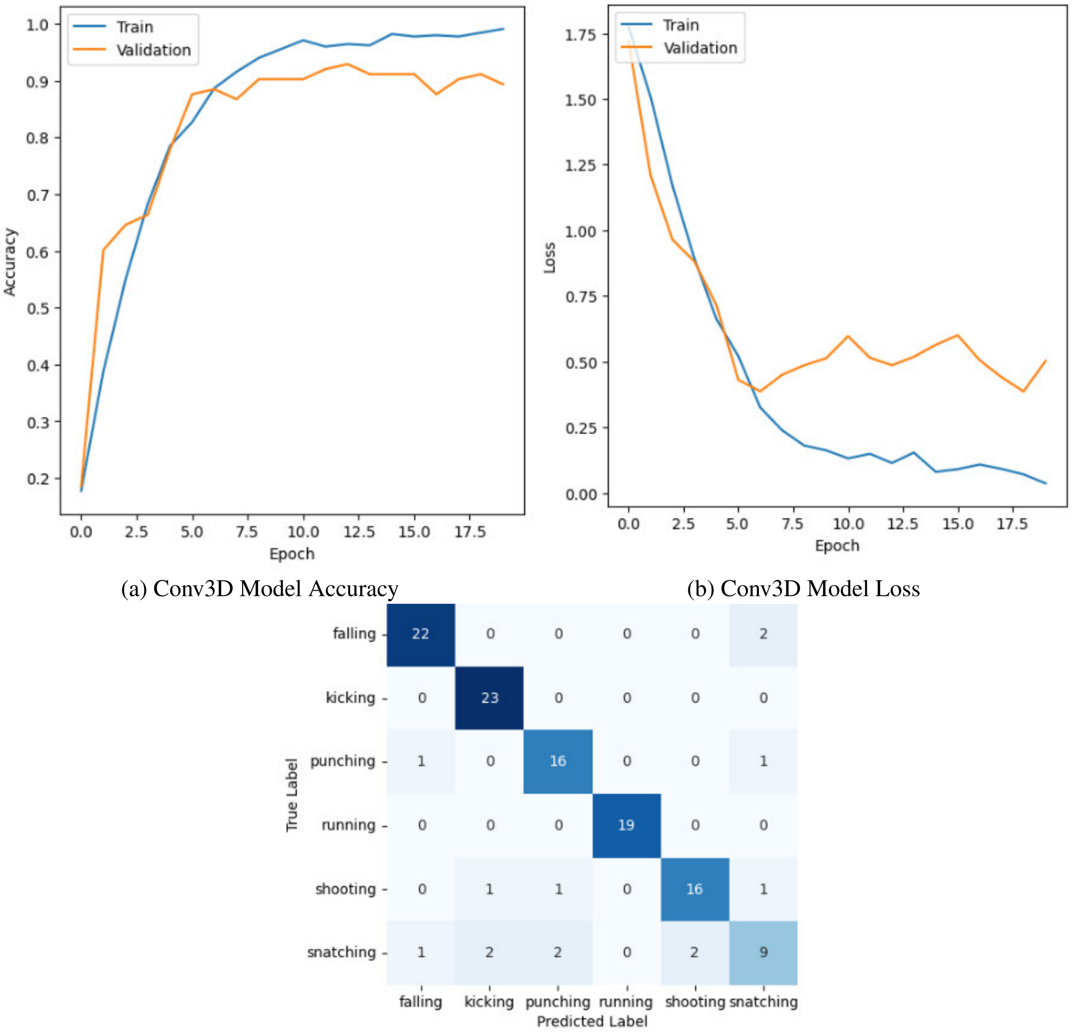
Model Loss'' (Figure 7b), and ''Conv3D Model Confusion Matrix'' (Figure 7c). These visualizations enhance the written description by offering graphical depictions of the model's accuracy, loss, and confusion matrix during several epochs.

In summary, these findings suggest that different models have different levels of efficacy in detecting and categorizing questionable human behaviours in movies. Among the models assessed, the Time Distributed CNN model demonstrates the highest level of accuracy.

C. MODEL PREDICTION

Model prediction entails utilizing a well-trained deep-learning model to produce accurate predictions or classifications on novel, unobserved data. Within this particular framework, the initial stage entails making predictions with the trained model using test data that has yet to be previously encountered. The test data is distinct from the data used for model training and is employed to assess the model's ability to generalize to novel, unseen situations. The given example in Figure 8a enumerates the prognostications generated by the model on a test video titled ''newfi32 - 6.avi.'' The model provides predictions for the probability of various activities taking place in the video, such as kicking, snatching, firing, punching, falling, and running. These predictions are accompanied by their respective confidence scores. As an example, the model accurately predicts with a confidence level of 99.12% that the action performed in the test video is ''kicking,'' while assigning significantly lower confidence ratings to alternative actions.

In the second stage shown in Figure 8b, the trained model is utilized to predict actions in a YouTube video, hence enhancing the accuracy of predictions. In the above scenario, the model accurately forecasts the action ''Kicking'' with a confidence score of 0.64% at the precise timestamp of 00:05. This forecast is generated by examining the frames of the YouTube video and categorizing them according to the acquired patterns and characteristics from the training data. The confidence score offered represents the model's degree of assurance in its forecast. In general, model prediction entails utilizing a well-trained model to analyze fresh data and make informed conclusions or classifications based on the patterns acquired from the training data. Automated analysis and categorization of data can be achieved using



(a) Conv3D Model Accuracy

(b) Conv3D Model Loss

(c) Conv3D Model Confusion Matrix

FIGURE 7. Conv3D Model Results Visualization.

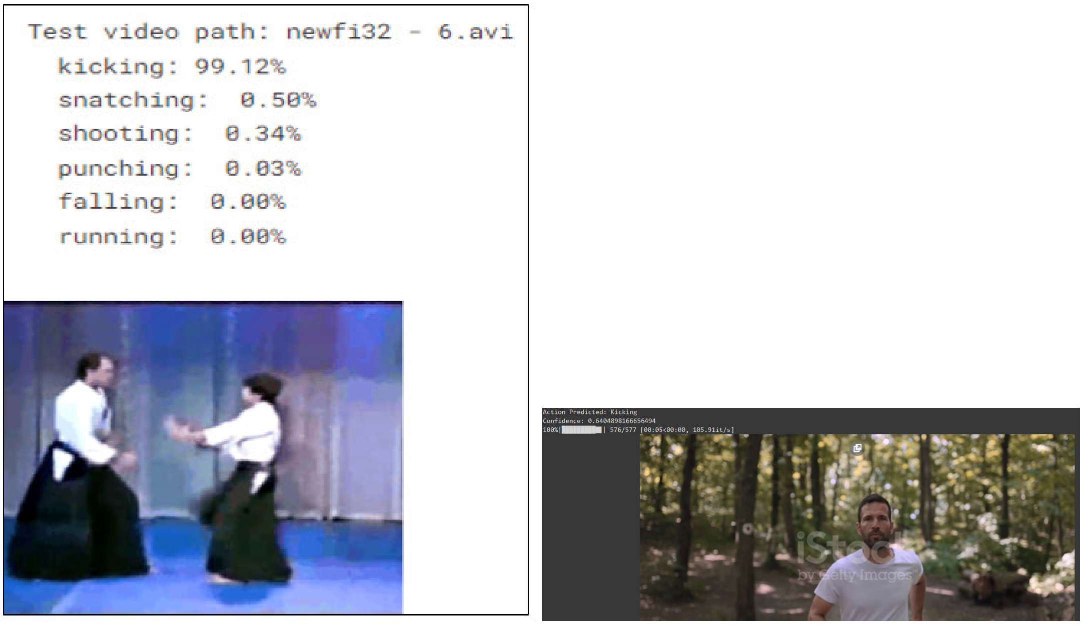
TABLE 2. Comparative Results for SHAR.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hybrid Model	84.51	84.89	83.10	84.72
Time Distributed CNN	90.14	90.78	90.14	90.14
Keras_GRU	83.80	87.10	84.10	84.10
Conv3D	88.23	88.20	88.20	88.20

this technology, making it highly useful in applications like video surveillance, object recognition, and natural language processing.

D. COMPARATIVE ANALYSIS

The comparison table, as depicted in Table 3, gives a complete assessment of both the proposed and existing research methodologies within a particular area. The cutting-edge methodology, as mentioned via Khan et al. citekhan2022human, encompasses various models, which include MLP (Multi-Layer Perceptron), CNN, LSTM, BiLSTM (Bidirectional LSTM), and CNN-LSTM, with every model's accuracy provided as a percentage. In contrast, the proposed method introduces novel models inclusive of the Hybrid model, Time disbursedTime, Keras_GRU, and Conv3D, each accompanied by using its corresponding accuracy score. Upon juxtaposing the modern and new methodologies, it becomes evident that the proposed strategies typically outperform the prevailing ones in phrases of accuracy. For instance, while the current method achieves a maximum accuracy of 76.50% using the CNN-LSTM model, the proposed Time allotted Timemodel demonstrates



```
Test video path: newfi32 - 6.avi
  kicking: 99.12%
  snatching:  0.50%
  shooting:  0.34%
  punching:  0.03%
  falling:  0.00%
  running:  0.00%
```

(a) Unseen Test Video Prediction

```
Action Predicted: Kicking
Confidence: 0.6404898166656494
100%|██████████| 576/577 [00:05<00:00, 105.91it/s]
```

(b) Unseen Youtube Video Prediction

FIGURE 8. Model Prediction.

TABLE 3. 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

a notably better accuracy of 90.14%. Further, the Conv3D model within the counselled method exhibits a sizable improvement in comparison to the contemporary techniques, reaching an accuracy of 88.23%. This evaluation underscores the capability of the presented models to supply advanced performance and outcomes within the specific domain below consideration, compared to the current methodologies. Comparative analyses of this nature play a pivotal role in helping researchers and practitioners identify the simplest processes for their precise tasks or applications.

VI. CONCLUSION AND FUTURE WORK

This study outlines a scientific method for detecting suspicious human activity through a series of crucial procedural steps. Our research started with the gathering of data from different resources, denoted as S1 and S2. We then meticulously refined this data by getting rid of inconsistencies, standardizing its format, and consolidating it right into a unified dataset. Rigorous image preprocessing ensued regarding normalization, scaling, and augmentation strategies to ensure image uniformity and enhance quality. Moreover, dataset annotation was undertaken to strengthen the precision of classifying and predicting suspicious human behaviour using supervised learning algorithms. To facilitate model training and evaluation, we partitioned the dataset into separate training and testing sets, ensuring unbiased assessment. CNNs have been employed to extract essential features from the preprocessed image data, which is important for future model implementation. To detect suspicious human activity, diverse deep-learning architectures, including the Hybrid LSTM model, time-dispersed CNN, Keras_GRU, and Conv3D, were explored. Each model leveraged the extracted CNN capabilities to gain insights into patterns and offer predictions. Our findings discovered that the proposed time-disbursed CNN model exeTimecuted an appreciably better accuracy rate of 90.14%, showcasing its efficacy in appropriately detecting suspicious human sports. Similarly, the Conv3D model in our proposed method exhibited full-size development as compared to current techniques, yielding an accuracy of 88.23%. In the end, we implemented the trained algorithms to forecast potentially suspicious human movements in real-world scenarios, which includes predicting the content of YouTube videos and scrutinizing surprising video footage. This sensible validation underscores the utility and relevance of our proposed methodology in bolstering surveillance and security systems by enhancing functionality to identify and mitigate potential threats.

Inside the realm of future research and advancement, several avenues present themselves for exploration and refinement. First of all, expanding the proposed technique to encompass a broader array of datasets, consisting of an extra range of suspicious activities and environmental factors, holds promise for boosting the model's efficacy and adaptableness. Additionally, delving into advanced deep learning architectures and methodologies, along with attention mechanisms and transformer models, can significantly enhance the accuracy and efficiency of activity recognition. Furthermore, integrating real-time facts streaming and processing competencies into the monitoring system ought to permit prompt identity and response to any suspicious activities, thereby improving typical security measures. Participating with domain experts and stakeholders is crucial to toughen the proposed technique's robustness and reliability. This collaboration can aid in refining the annotation process of the research and validating the model's predictions in real-time scenarios. Furthermore, considering the ethical considerations and privacy ramifications related to surveillance systems, forthcoming research needs to prioritize the development of obvious and responsible frameworks for statistics series, storage, and utilization. This necessitates the implementation of sturdy records protection measures and adherence to pertinent regulatory requirements to protect individuals' privacy rights while retaining robust security protocols. By way of addressing these important regions of challenge, future research endeavours can contribute to the evolution of surveillance systems that are both ethically sound and operationally effective.

REFERENCES

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

[2] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2662–2666.

[3] C. V. Amrutha, C. Jyotsna, and J. Amudha, "Deep learning approach for suspicious activity detection from surveillance video," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Mar. 2020, pp. 335–339.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[5] J. Wei, J. Zhao, Y. Zhao, and Z. Zhao, "Unsupervised anomaly detection for traffic surveillance based on background modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 129–136.

[6] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A. E. Hassanien, and H. M. Pandey, "An optimized dense convolutional neural network model for disease recognition and classification in corn leaf," *Comput. Electron. Agricult.*, vol. 175, Aug. 2020, Art. no. 105456.

[7] R. Teja, R. Nayar, and S. Indu, "Object tracking and suspicious activity identification during occlusion," *Int. J. Comput. Appl.*, vol. 179, no. 11, pp. 29–34, Jan. 2018.

[8] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in LSTMs for activity detection and early detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1942–1950.

[9] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[10] S. Ghazal, U. S. Khan, M. Mubasher Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," *IET Image Process.*, vol. 13, no. 13, pp. 2572–2578, Nov. 2019.

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

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

[13] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," *IET Comput. Vis.*, vol. 12, no. 1, pp. 16–26, Feb. 2018.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[16] J. Li, R. Wu, J. Zhao, and Y. Ma, "Convolutional neural networks (CNN) for indoor human activity recognition using ubisense system," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 2068–2072.

[17] L. Anishchenko, "Machine learning in video surveillance for fall detection," in *Proc. Ural Symp. Biomed. Eng., Radioelectron. Inf. Technol. (USBEREIT)*, May 2018, pp. 99–102.

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

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

[20] U. M. Butt, S. Letchmunan, F. Hafinaz, S. Zia, and A. Baqir, "Detecting video surveillance using VGG19 convolutional neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, 2020.

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

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

[23] M. Qasim Gandapur and E. Verdú, "ConvGRU-CNN: Spatiotemporal deep learning for real-time anomaly detection in video surveillance system," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 8, no. 4, p. 88, 2023.

[24] R. Rajeswari, "Anomalous human activity recognition from video sequences using brisk features and convolutional neural networks," *Galaxy Int. Interdiscipl. Res. J.*, vol. 10, no. 2, pp. 269–280, 2022.

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

MONJI MOHAMED ZAIDI was born in Kairouan, Tunisia, in January 1982. He received the B.S. degree in electrical engineering for automation and control process from the National Engineering School of Sfax, Tunisia, in 2005, and the M.S. degree in nanostructures, devices and micro-electronics systems (NDMS) and the Ph.D. degree in electronics from the University of Monastir, Tunisia, in 2007 and 2011, respectively. He has been an Assistant Professor with the College of Engineering, King Khalid University, Saudi Arabia, since September 2014. He has been an Assistant Professor with the Higher Institute of Computer Science and Mathematics of Monastir, Tunisia, since 2013. His research interests include the management of wireless technologies and hardware-software co-design for rapid prototyping in telecommunications.

GABRIEL AVELINO SAMPEDRO (Member, IEEE) is currently with the School of Management and Information Technology, De La Salle—College of Saint Benilde, Philippines. He has worked with the Faculty of Information and Communication Studies, University of the Philippines Open University, Laguna, Calabarzon, Philippines. His current research interests include the Internet of Things and artificial intelligence.

AHMAD ALMADHOR received the B.S.E. degree in computer science from Aljouf University (formerly Aljouf College), Al Jowf, Saudi Arabia, in 2005, the M.E. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2010, and the Ph.D. degree in electrical and computer engineering from the University of Denver, Denver, CO, USA, in 2019. From 2006 to 2008, he was a Teaching Assistant, the College of Sciences Manager, and then a Lecturer, from 2011 to 2012, Aljouf University. Then, he became a Senior Graduate Assistant and a Tutor Advisor with the University of Denver, in 2013 and 2019. He is currently an Assistant Professor in CEN and VD with the Computer and Information Science College, Jouf University, Saudi Arabia. His research interests include AI, blockchain, networks, smart and microgrid cyber security, integration, image processing, video surveillance systems, PV, EV, machine, and deep learning. His awards and honors include the Aljouf University Scholarship (Royal Embassy of Saudi Arabia in D.C.) and Aljouf's Governor Award for Excellency.

SHTWAI ALSUBAI received the bachelor's degree in information systems from King Saud University, Saudi Arabia, in 2008, the master's degree in computer science from CLU, USA, in 2011, and the Ph.D. degree from The University of Sheffield, U.K., in 2018. He is currently an Assistant Professor in computer science with Prince Sattam Bin Abdulaziz University. His research interests include XML, XML query processing, XML query optimization, machine learning, and natural language processing.

ABDULLAH AL HEJAILI received the bachelor's degree in computer science from the Tabuk Teachers College, Saudi Arabia, in 2007, and the master's degree in computer science from CLU, USA, in 2011. He is currently pursuing the Ph.D. degree with the Informatics School, University of Sussex. He is currently a Lecturer in computer science with the University of Tabuk. His research interests include technology-enhanced learning, image processing, virtual and augmented reality, motion capture, and education applications.

MICHAL GREGUS is currently with the Faculty of Management, Comenius University in Bratislava. His research interests include but is not limited to artificial intelligence, data management, data analysis, and data science.

SIDRA ABBAS (Graduate Student Member, IEEE) received the B.S. degree from the Department of Computer Science, COMSATS University Islamabad, Pakistan. Her research interests include but are not limited to computer forensics, machine learning, criminal profiling, software watermarking, intelligent systems, and data privacy protection.