The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

The Real-time fitness action recognition using LSTM/GRU techniques

Azhee Wria Muhamad[1,2], Aree Ali Mohammed[3]

[1] Computer Science Department, College of Basic Education, University of Suleimani, Suleimani, Iraq

[2] Information Technology Department, Tishk International University, Suleimani, Iraq

[3] Computer Science Department, College of Science, University of Suleimani, Suleimani, Iraq

Email: azhee.muhamad@univsul.edu.iq[1], aree.ali@univsul.edu.iq[3]

Abstract:

Human action recognition in fitness refers to the ability of technology to identify and track human movements during the exercise or physical activity. This may include recognizing specific exercises, such as Squats or push-ups, as well as tracking the overall movement patterns and providing feedback on forms and techniques. This technology is often used in fitness apps or wearable devices to help individuals improve their workouts and prevent injury. Using Artificial Intelligence (AI) to recognize human actions in fitness is a great idea. With the help of Artificial Intelligence, fitness enthusiasts can get accurate and real-time feedback on their workout routines, which can help them improve their performance and achieve their fitness goals faster. However, it is important to ensure that the data collected by these devices are secure and not misused in any way. Overall, it is believed that AI has the potential to revolutionize the way we approach fitness and lead to a healthier lifestyle for many people. Recurrent networks consist of distributed parameters that are present in each layer of the network. This study introduces dual models for recognizing many actions, one using LSTM and the other using GRU. An experiment was carried out on the EUCF Sport action dataset to compare the accuracy rates of the two models. The results indicate that both the LSTM and GRU models had significantly better accuracy rates than other state-of-the-art action recognition models, with recorded accuracies of 99.78% and 98.53% respectively on EUCF Sports dataset.

Keywords: Human action recognition, LSTM, GRU, Accuracy, and UCF dataset.

الملخص:

يتضمن التعرّف على أفعال الإنسان في مجال اللياقة البدنية استخدام تقنيات تقوم بتحديد وتتبع حركات الجسم أثناء التمارين أو الأنشطة البدنية. وتُستخدم هذه التقنيات على نطاق واسع في تطبيقات اللياقة البدنية والأجهزة القابلة للارتداء بهدف تحسين جودة التمارين والحد من الحركات الخاطئة. تتكوّن الشبكات العصبية المتكررة من معاملات موزعة تتكرر عبر طبقات الشبكة المختلفة.تقدّم هذه الدراسة نموذجين مزدوجين للتعرّف على مجموعة من الأفعال؛ يعتمد الأول على شبكة الذاكرة طويلة وقصيرة المدى(LSTM) ، بينما يعتمد الثاني على وحدة البوابات المتكررة .(GRU) أُجريت تجربة على مجموعة بيانات **EUCF Sports**(الموسّعة من جامعة وسط فلوريدا) لمقارنة معدلات الدقة بين النموذجين. ويتمثل الهدف الرئيس في تحقيق دقة أعلى باستخدام أسلوب الذاكرة طويلة وقصيرة المدى (LSTM) الذي استُخدم لتحسين نموذج الشبكات العصبية المتكررة(RNN). تشير النتائج إلى أن كلا النموذجين، LSTMوGRU، حققا معدلات دقة أعلى بشكل ملحوظ مقارنةً بنماذج التعرّف على الأفعال المتقدمة الأخرى، حيث سُجلت دقة قدرها **99.78%**لنموذج LSTM و **98.53%**لنموذج GRU على مجموعة بيانات **EUCF .Sports**

الكلمات المفتاحية: الدقّة، وحدة البوابات المتكرّرة (GRU)، التعرّف على أفعال الإنسان، الذاكرة طويلة وقصيرة المدى (LSTM).

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

پوختە:

ناسینەوەی کردارەکانی مرۆڤ لە بواری فیتنەسدا بریتییە لە بەکارهێنانی تەکنەلۆجیا بۆ ناسینەوە و شوێنکەوتنی جوڵە و هەڵسوکەوتەکانی مرۆڤ لە کاتی ڕاهێنان یان چالاکیی جەستەیی. ئەم تەکنەلۆجیایە بەزۆری لە ئەپلیکەیشنەکانی فیتنەس و ئامێرە لەبەرکراوەکاندا بەکاردێت بۆ باشترکردنی ڕاهێنان و ڕێگریکردن لە جوڵەی هەڵە. تۆڕە دەماری دووبارەبووەکان (Recurrent Neural Networks)یێکهاتوون لە پارامێتەرە هاوبەشەکان کە لە هەموو چینەکانی تۆڕەکەدا دووبارە دەردەکەون.ئەم توێژینەوەیە دوو مۆدێلی جیاواز پێشکەش دەکات بۆ ناسینەوەی ژمارەیەک کردار؛ یەکێکیان پشت بە تۆڕی بیرەوەری درێژ و کورتماوە (LSTM)دەبەستێت و یەکێکی دیکەیان پشت بە یەکەی دەروازەدارە دووبارەبوو (GRU). تاقیکردنەوەیەک ئەنجامدرا لەسەر کۆمەڵەداتای EUCF Sports (وەشانێکی فراوانکراوی زانکۆی ناوەندی فلۆریدا) بۆ بەراوردکردنی ڕێژەی وردبینی ئەم دوو مۆدێلە. ئامانجی سەرەکی ئەم کارە گەیشتنە بە وردبینییەکی باشتر بە بەکارهێنانی شێوازی LSTM-ە، کە بۆ باشترکردنی مۆدێلی RNNبەکارهاتووە. ئەنجامەکان پیشان دەدەن کە هەردوو مۆدێلی LSTM و GRU بە شێوەیەکی بەرچاو ڕێژەی وردبینییەکی بەرزتریان هەیە بەراورد بە مۆدێلە پێشکەوتووەکانی دیکەی ناسینەوەی کردار، بە وردبینیی تۆمارکراوی 99.78%بۆ LSTM و 98.53%بۆ GRU لەسەر کۆمەڵەداتای EUCF Sports.

کلیلە وشە: وردبینی، یەکەی دەروازەدارە دووبارەبوو ، ناسینەوەی کردارەکانی مرۆڤ، بیرەوەری درێژ و کورتماوە.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Introduction

Recognition of human action identification is a field of computer vision (CV) and machine learning (ML) that involves identifying and classifying human actions from visual data. Recognising human action's purpose is to accurately recognize the actions performed by humans in a given video or image sequence. This involves detecting and tracking human body parts, extracting relevant features from the video data, and then classifying the actions performed based on those features. The majority of approaches used by humans to recognize actions struggle to identify actions in lengthy videos that consist of multiple scenes and actions. There are multiple types of human activities, such as interactions between humans and interactions between humans and objects.

There are several challenges in human action recognition, including variability in human appearance, motion, and scene context, as well as occlusions and other visual distractions. To overcome these challenges, various approaches have been developed, including deep learning-based methods, feature-based methods, and hybrid methods [1]. Human action recognition has numerous applications, such as in video surveillance for detecting suspicious behaviours, in fitness analysis for tracking and analysing athlete movements, in robotics for controlling robots based on human actions, and in entertainment for creating realistic virtual characters and special effects [2].

The process of vision-based human activity recognition involves several steps. First, the video data are captured using cameras or other imaging devices. Next, the video frames are pre-processed to remove noise and enhance the features of interest. Then, feature extraction techniques are used to extract relevant information from the video frames. Once the features have been extracted, machine learning algorithms are used to classify the activities being performed in the video [3]. These algorithms can be trained using labelled data sets that contain examples of different fitness activities as showing Figure 1. Some common techniques used for feature extraction include motion analysis, shape analysis, and appearance-based methods. Machine learning algorithms used for classification include support vector machines (SVMs) [4], decision trees, and neural networks. Overall, vision-based human activity recognition has the potential to revolutionize many industries by providing automated monitoring and analysis of human behavior.

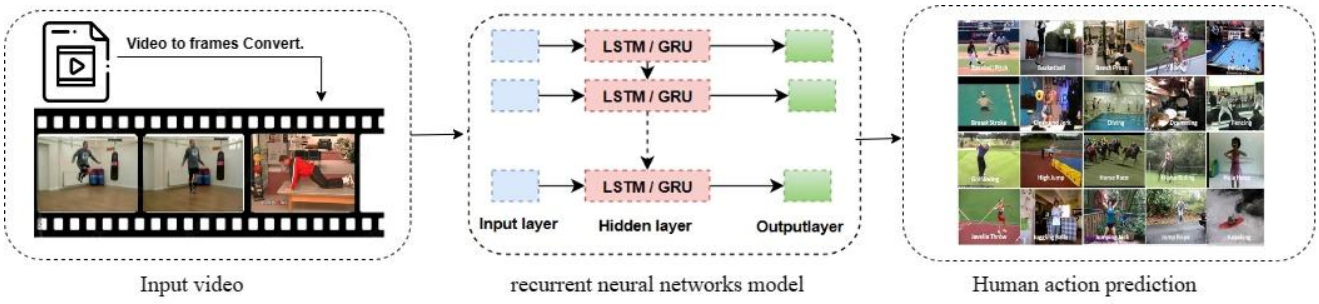


Figure 1. A description of the model for the prediction of human actions

Recently, deep learning has shown promising results in the field of computer vision. Deep learning models consist of multiple layers of artificial neurons that process and transform input data, gradually learning higher-level features and representations of the data. The number of layers and the complexity of the model architecture can vary depending on the task and the amount of data

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

available. These models consist of multiple processing layers such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Figure 2 displays the four categories of activities, which include gestures, actions, interactions, and group activities. These activities are classified based on their complexity and duration [5].

- A gesture: refers to a fundamental motion made by different parts of the human body that conveys a certain message. Examples of gestures include 'hand shaking', and 'leg waving'. Typically, a gesture lasts only a brief period of time.
- An action is a sequence of physical movements performed by a single a person. Human activities include strolling, jogging, sprinting, and punching.
- Interaction: involves two actors, one of which is always a human, while the other could be either a human or an object. Human-human interactions include fighting, handshaking, and hugging, while human-object interactions include using a treadmill and a computer.
- Group activity: The most complicated kind of activity is a group activity, which combines gestures, actions, and interactions. It involves one or more things and more than two people. A game between two teams and a group gathering are two examples of group activities.

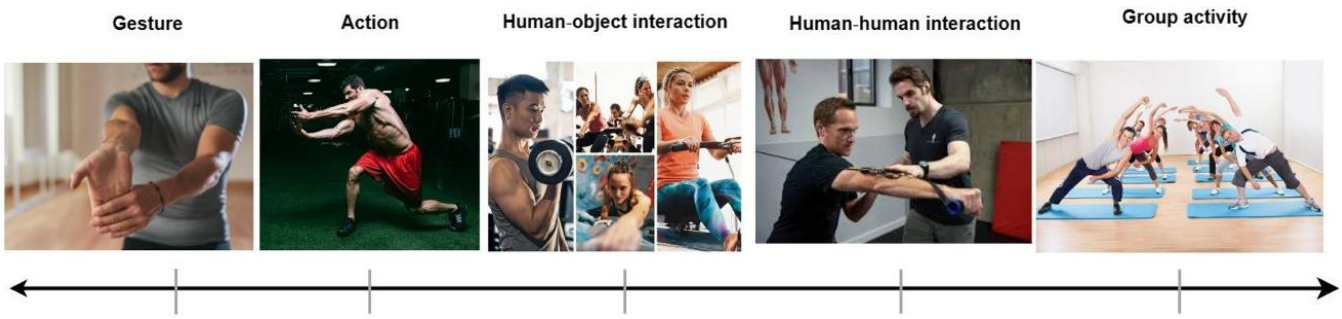


Figure 2. Classification of various tiers of actions.

However, the objective of human action recognition in fitness is to accurately identify and classify different physical activities performed by individuals during their workout sessions. This technology can be used to monitor and track progress, provide feedback on form and technique, and personalize workout plans based on individual needs and goals. It can also be used in virtual coaching applications to provide real-time guidance and motivation to users. Overall, the objective is to enhance the effectiveness and efficiency of fitness training through advanced technology.

This research introduces three significant point.

1. These models make it possible to learn both forwards and backwards about the order of events from the parts of a video frame.
2. The models have an extraordinary capability for learning sequences and adapting their characteristics from one frame to the next.
3. Making the RNN design model work better for feature extraction, several pre-processing techniques are applied. Additionally, some techniques are utilized to normalize the features before they are fed into the model. Because of these attributes, the proposed approach is better suited for identifying actions in videos.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

The remaining sections of the article are organized as follows: Section two reviews relevant works; Section three explains the methodology; Section four reports testing findings; Section five evaluates the method and compares it to other modern methods; and Section six discusses future research directions.

Related Works

In recent years, human action recognition has attracted considerable interest and been the subject of extensive research. Researchers have been endeavouring to identify human movements in images and videos since the 1980s. One of the main strategies for action identification that academics have been exploring is comparable to how the human visual systems operate. The human visual system is able to receive a number of observations about the motion and form of the human body in a short amount of time when it is working at a low level. Then, these data are transmitted to the intermediate human perception system in order to further recognise the categorization of these observations, such as jogging, walking, and running [6]. Researchers have put in a lot of work over the past few decades to get a computer-based recognition system to work as well as a human system. Still, the researchers are a long way from the level of the visual system in humans. However, there are several difficulties and problems associated with HAR, including environmental complexity, the non-rigid shapes of people and objects, as well as another problem associated with human recognition, are characterised by their flexibility.

2.1 Handcrafted representation-based technique

Conventional action recognition methods entail the construction of an action representation. This technique has become popular within the Human Activity Recognition (HAR) community and has produced remarkable results on a number of well-known public datasets. Classification is then achieved by training a rudimentary classifier, such as a Support Vector Machine (SVM) [7], using the extracted features from the sequence of videos. This strategy incorporates space-time and appearance-based methods.

Space-Time-Based Approaches

Space-Time-Based Approaches refer to the use of spatiotemporal information to identify and classify human actions in videos. These approaches typically extract motion features from video sequences and use them to model the dynamics of human actions over time and space. There are several techniques and algorithms used in Space-Time-Based Approaches for human action recognition [8]. One of the most popular techniques is the use of optical flow, which describes the motion of objects between two consecutive frames of a video.

There are four primary elements to space-time-based strategies, including a space-time interest point (STIP) detector, a feature descriptor, a vocabulary builder, and a classifier [9]. The STIP detectors can be classified as either dense or sparse detectors. Detecting interest point s in video content can be achieved using two types of detectors: dense detectors (V-FAST, Hessian detector, and dense sampling) that cover the video content extensively, and sparse detectors, like the cuboid detector, Harris3D, and Spatial-Temporal Implicit Shape Model (STISM), which focus on a small

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

(sparse) part of the information [10, 11]. Multiple STIP detectors have been created by various researchers.

Feature descriptors may also be divided into local and global descriptors. Local descriptors use data that is peculiar to a location, such texture, colour, and posture (such as cuboid descriptor, Enhanced Speeded-Up Robust Features (ESURF), and N-jet). Global descriptors, on the other hand, make use of data that is more evenly distributed across the video, such as fluctuations in speed, phase, and lighting. Either the bag-of-words model or the state-space model is used in the techniques for expanding vocabulary or acquiring data [8]. As shown in Figure 3, either a supervised or unsupervised classifier is used to categorise the data.

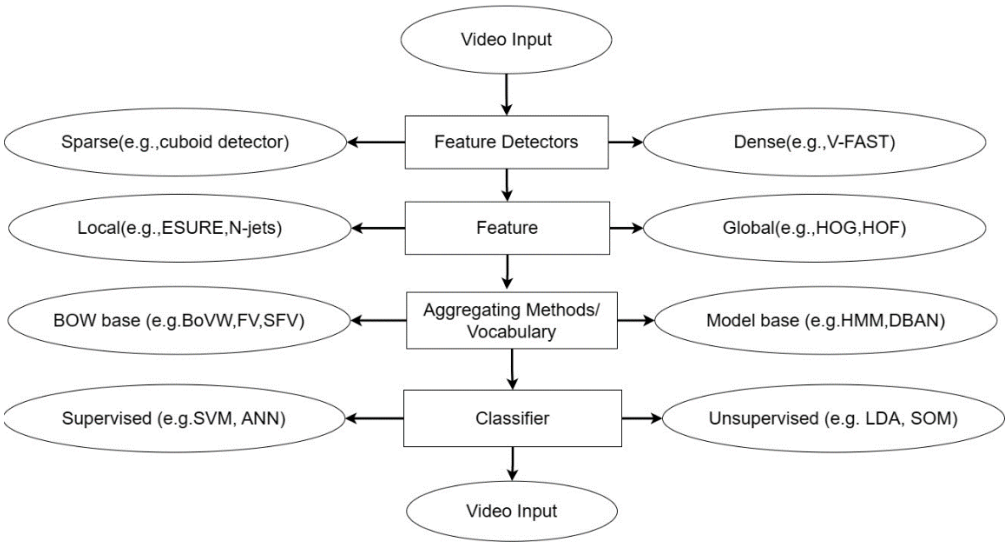


Figure 3. Space-time-based methods have different parts. [8].

Appearance-Based Approaches

This section explores two types of approaches for recognizing actions: those that use 2D (XY) and 3D (XYZ) depth images and rely on shape, motion, or a combination of both to extract features. The 2D shape-based approaches extract features based on shape and contour to represent actions [12], while the motion-based approaches utilize optical flow or similar techniques to extract features that represent motion during action. There are methods that incorporate both shape and motion features to represent and recognize actions [1].

3D-based methods require the creation of a representation of the human body for actions, which can be constructed using a variety of shapes including cylinders, ellipsoids, visual hulls derived from silhouettes, and surface geometry. This category includes techniques such as 3D optical flow, shape histogram, motion history volume, and body skeleton. Shape-based approaches in HAR involve using the shape of the human body as the primary feature for recognizing actions. These approaches rely on constructing a model of the human body and analysing its shape or changes over time to recognize different actions. One commonly used technique is the use of point clouds, which represent the shape of the human body using a collection of 3D points [13]. These points can be used to calculate various features, such as the curvature of the body or the distribution of points across different body parts.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

2.2 Learning-based Action Recognition Approach

3D-based methods require the creation of a representation of the human body for actions, which can be constructed using a variety of shapes including cylinders, ellipsoids, visual hulls derived from silhouettes, and surface geometry. This category includes techniques such as 3D optical flow, shape histogram, motion history volume, and body skeleton. as well as deep learning-based models. These approaches can be categorized both groups: non-deep learning-based and deep learning-based approaches [14], as illustrated in figure 4.

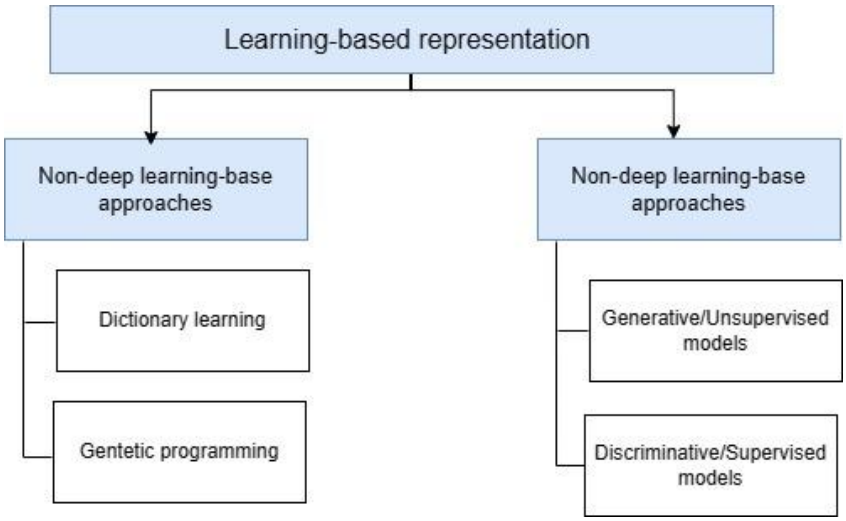


Figure 4. Methods for representing actions based on learning [14]

2.2.1 Approaches Not Based on Deep Learning

These techniques, as explained here, are based on genetic programming and dictionary learning. Dictionary learning is a commonly used method for representation learning, wherein the primary focus is on learning sparse representations of input data through the use of a dictionary of basic atoms. This method has been used in many computers vision tasks, such as classifying images and recognising actions [15]. This technique creates a concise representation of the input data by using a linear combination of basis dictionary atoms. The process involves unsupervised learning to simultaneously learn the dictionary and classifier. This method is comparable to the visual Bag of Words model, which generates global data representations. One example presented in [16] is a weakly-supervised cross-domain dictionary learning approach that adapts knowledge from one action dataset to another. The method uses reconstructive, discriminating, and domain-adaptive dictionary pairs and corresponding classifier parameters without prior information. Spatial-temporal feature-based methods are popular for activity classification, but the use of over-complete dictionaries is more interesting as they produce even more compact representations.

Genetic programming is an evolutionary technique that mimics natural evolution and can be used to solve problems without knowing the solutions beforehand. In the field of human activity recognition, genetic programming can determine the optimal sequence of primitive operations to optimise recognition performance. A recent example of this approach is presented in [17], which focuses on using genetic programming for action recognition.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

2.2.2 Deep Learning-Based Approaches

In recent years, feature learning has become popular in a wide variety of computer vision applications, including pedestrian detection, image classification [18], vision-based anomaly detection [17], etc. Several learning-based approaches to action recognition involve converting pixels into action classes through end-to-end learning. The feature-learning-based HAR task is constructed on deep learning. The multilayer convolutional neural network with long short-term memory (LSTM) combined in a deep neural network (LSTM) could automatically extract action characteristics and categorize them based on some of the model parameters. Recurrent neural networks (RNN) such as the LSTM are particularly good at processing temporal sequences. A successful case study of frame classification is what inspired action recognition in video sequences [19].

Deep bidirectional LSTM and convolutional neural network (CNN) networks are used in a technique for video analysis [18]. This method is meant to improve accuracy. S. K. Choudhury et al.2017 used a LSTM structure to improve the speed, pre-processing ability on the video, and flexibility of video recognition. Having achieved success in video frame recognition has renewed interest in deep learning in video identification. To deal with input flow, using memory blocks with gating architecture and LSTMs to address the gradient vanishing problem and improve the extraction of long-term characteristics, a series of memory blocks with gating architecture is used. There is a suggestion that LSTM neural networks capture more active and expressive temporal representations. When training the LSTM model simultaneously on the demanding benchmark databases, UCF Sports and EUCF Sports, identification accuracy is best than when using handcrafted features.

The Methodology

RNNs are artificial neural networks that have connections between units that form directed cycles. The RNN presented in this study was improved by modifying LSTM and GRU models in order to improve accuracy in recognizing human actions on the UCF Sports dataset. Preprocessing algorithms, in order to improve the quality of the dataset, several methods were used, such as converting colours to grayscale, equalizing histograms, applying filters, and normalizing the data. To prevent overfitting, the researchers increased the amount of training data and used augmenting data techniques as shown in Figure 5. The analysis concentrated on two types of recurrent neural networks (RNNs): LSTM and GRU. The efficacy of these two networks was compared to previous models that used LSTM and GRU designs.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

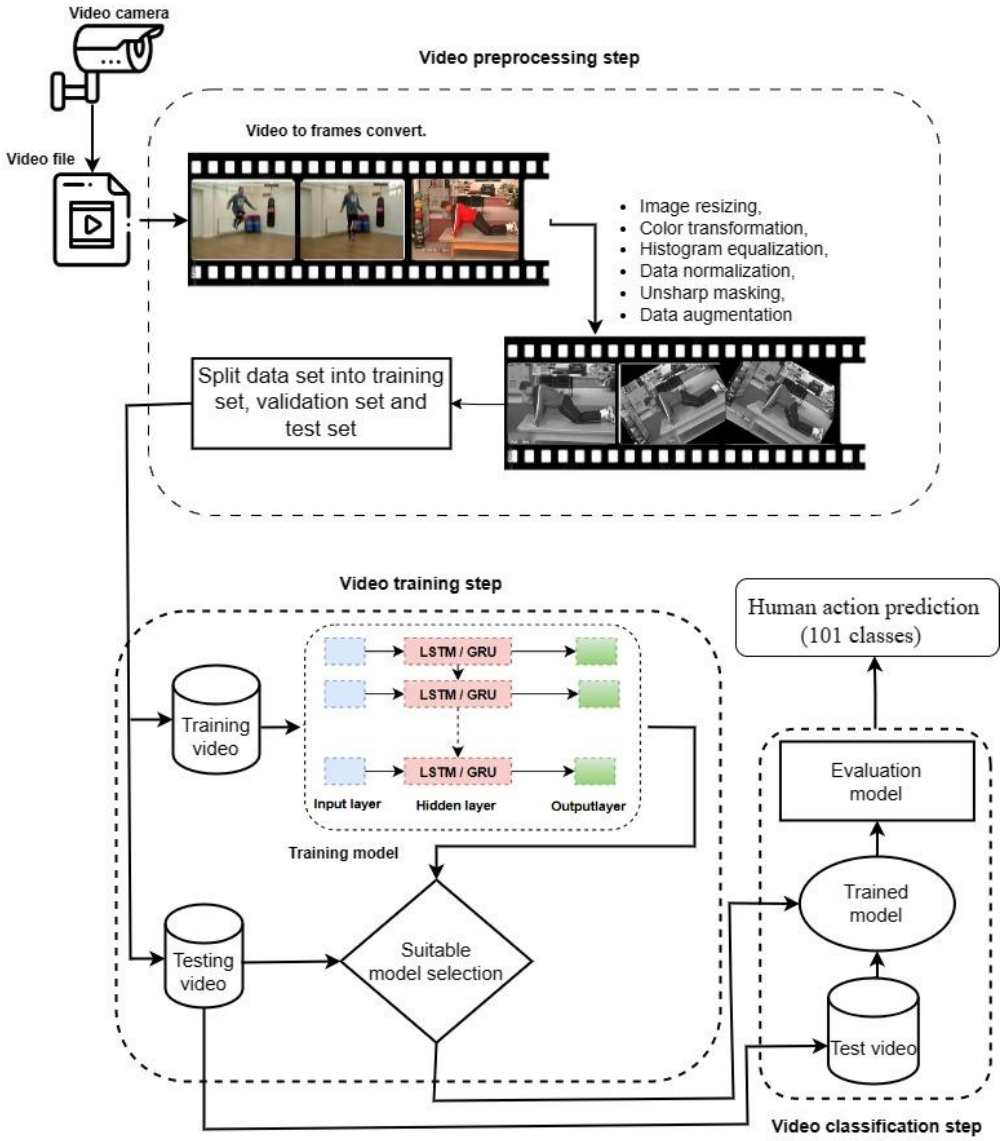


Figure 5. Block representation of the model's architecture

3.1 The Long Short-Term Memory (LSTM)

Since the traditional RNN system is not very good at predicting, thus the LSTM architecture is used instead. LSTMs are great for remembering data for long and brief periods. The LSTM blocks have either 3 or 4 gates which employ the logistic function to ascertain values from 0-1. This value decides how much information enters or leaves the memory. The model has an input gate to control the flow of data, a forget gate to control how long memory is kept, and an output gate to control how much data is used to make the block's output [20]. The modified RNN model uses a higher level of LSTM architecture, with four layers compared to just one in the traditional model, as showing in figure 6.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

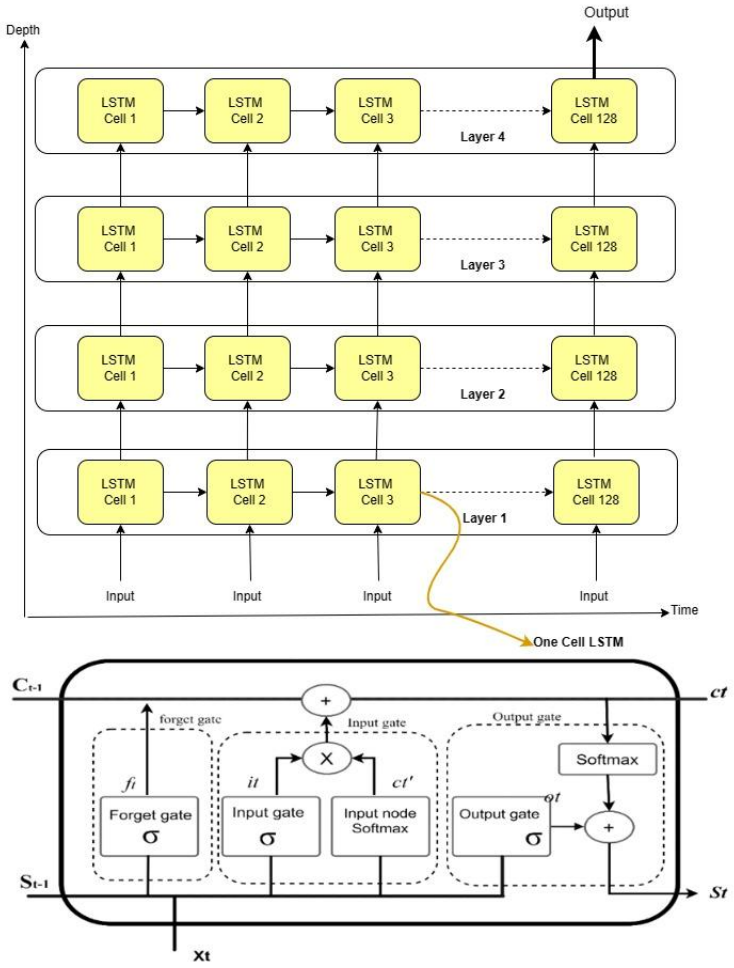


Figure 6. Block representation of the LSTM model.

The LSTM network model has been shown to work better than the standard model. It is made up of the following parts:

1- The LSTM model has more layers than it did before. Long Short-Term Memory algorithms are good at recognizing temporal information in sequential data, but they can be affected by the gradient vanishing problem. By adding more LSTM layers to the model, the neural networks can be depended, making it more accurately considered a deep learning model. This type of depth has been associated with successful predictions when it comes to the human action in videos. The LSTM model is better than basic LSTMs because it has memory cells that help pull features out of sequential data. The model in the study runs the raw data through four layers of LSTM in order to figure out how the sequence data relates to time.

2. Adding more units to every layer of this model can be compared to describing the hidden state or output. Memory cells exist in this model's concealed state. The number of units shows how many neurons are connected to the layer that has the secret state vector and input.

3. By changing the parameters or settings of neural networks, various optimization strategies may be utilized to reduce mistakes. This optimization process involves narrowing the gap between predicted and actual outcomes. To evaluate the consequences of using Adam and SGD optimizers, an LSTM model was utilized.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Stochastic Gradient Descent (SGD), one of the most popular ways to improve neural networks, is a gradient descent method that changes the values of its arguments with each run to reduce mistakes. Furthermore, the Adam Optimizer combines the advantages of Adaptive Gradient (AdaGrad) and Root Mean Square Propagation (RMSProp) by utilizing both the primary and next moments of the gradient. This approach updates the squared gradient and exponential moving average of first- and second-moment gradient estimations. Adam optimizer outlines optimizations processes. [21].

3.2 The Gated Recurrent Unit

The GRU network is a popular type of RNN among many variations. It relies on the previous hidden state for activation each time. However, RNNs are not easy to train due to the disappearing gradient effect. Nevertheless, some variants like the GRU have been shown to perform well in tasks that require long-term data retention, such as generating captions for images or videos. Unlike the LSTM design, where the input and forget gates are separate, the GRU has combined them into one update gate. The LSTM architecture uses three gating signals, while the GRU only uses two. The GRU design is seen in Figure 7 and consists of a reset gate and an update gate, with the former controlling the speed at which data from the previous state reaches the current one.

The GRU model employs several parameters: z_t represents the update gate, x_t represents the input, r_t represents the reset gate, and y_t represents the output. The previous state, h_{t-1}, is multiplied by appropriate weights, and z_t represents the update gate, which determines the degree to which the unit alters its activation. When r_t gets close to zero, the reset gate ignores the state it has already calculated and acts like it is handling the first symbol. To get z_t and r_t, a sigmoid activation is used. The reset gate decides how much of the data from the last time step to keep and how much to throw away. The following equation shows information about the forward path at time step t. [22]:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (7)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (8)$$
$$h'_t = Softmax(W x_t + r_t * h_{t-1}) \quad (9)$$
$$h_t = (1 - z_t) h_{t-1} + z_t h'_t \quad (10)$$
$$y_t = \sigma(W_o h_t) \quad (11)$$

The GRU network model, as compared to the traditional model, is comprised of the following components:

1. More layers have been added to the GRU model. This model is very suitable for pulling out features from sequential data. The model uses two layers of GRU to pull timing traits out of the sequence data that is given as input.
2. Each stage of the model incorporates additional GRU units. The hidden state of the model is made up of memory cells, and in the layer where concealed states and inputs are combined, the number of units specifies how many neurons are connected.
3. The GRU algorithm has used a variety of optimizers. This model is advantageous for preserving important information over long sequences for it has less gates, removes unnecessary data, and stores information dependencies in its hidden state. The optimizer algorithms used are Adam and

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Stochastic Gradient Descent, with a learning rate, to improve first-order optimization. The stochastic gradient descent-based algorithm allows for the dynamic updating of relevant parameters.

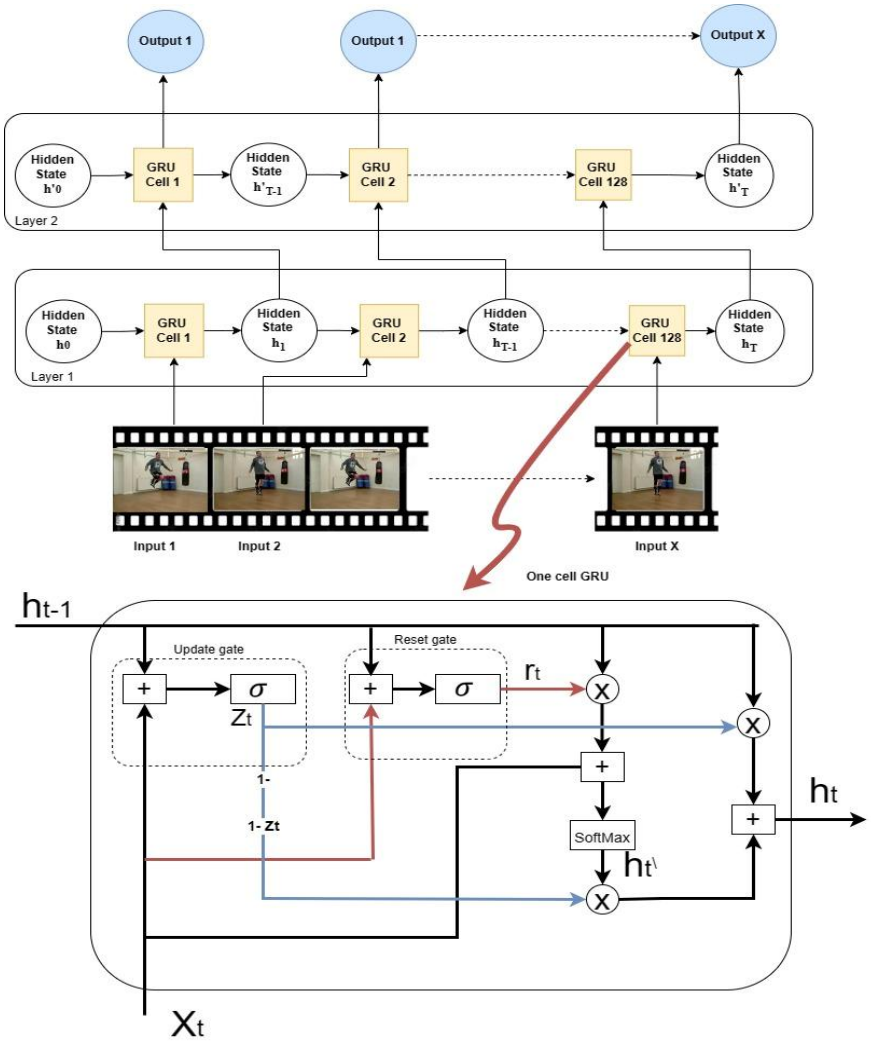


Figure 7. Block representation of the GRU model.

3.3 Dataset preparation

The study provides an excellent overview of those early datasets. Newer datasets, such as UCF, Kinetics and Hollywood, generally, contain unconstrained videos that simulate real-world scenarios. Several datasets are available online that can be utilized to pretrain models, and among them, there is a UCF dataset that is appropriate for human action. Pre-trained models may be trained using a variety of internet-accessible datasets, whereas human activity may be best suited to a UCF dataset.

3.3.1 The UCF101

The UCF101 database has 13,320 YouTube video clips with a present frame rate of 25 fps and a size of 320 x 240 pixels. It is made up of 101 distinct human activity classes. For each activity class, the video clips are divided into 25 groups, each including four to seven portions. Under the existing literature for UCF101, train and test separations are employed for recognition of act in order to

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

prevent video clips from the same movie from being used for both training and testing. But given that modern algorithms may attain accuracy levels of 95% or greater, it's possible that the datasets don't accurately reflect actual data [24]. Figure 8 provides an illustration of this.

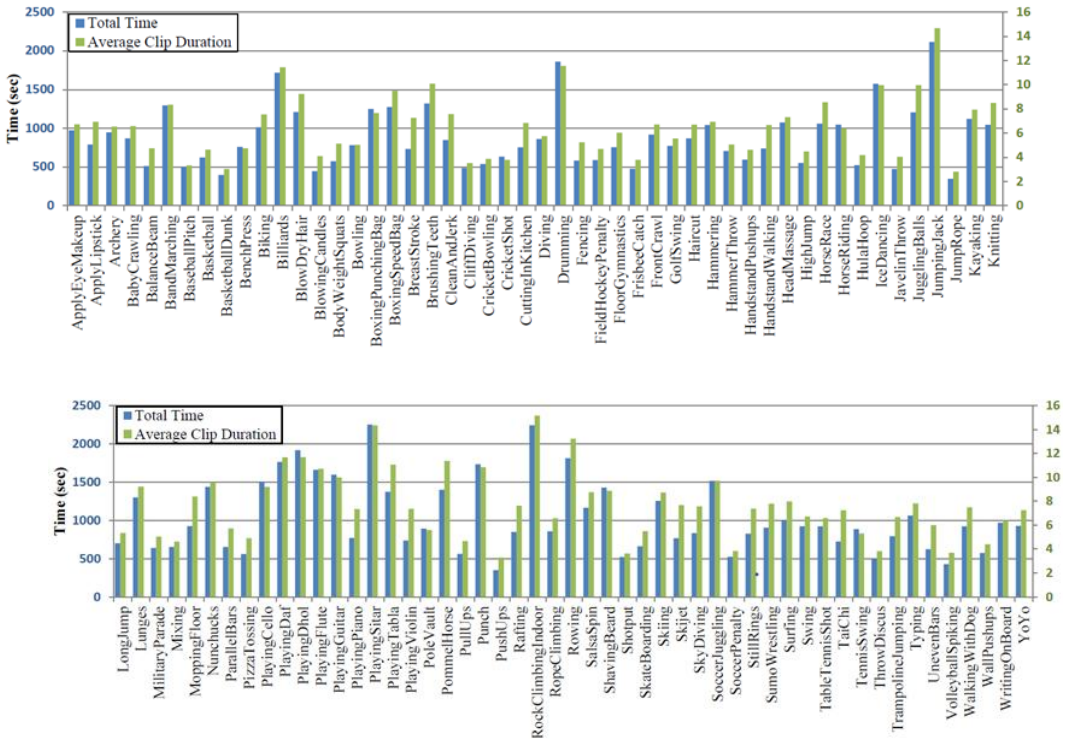


Figure 8: UCF101 has 101 actions, the total time of video clips for each human action class is blue. The average length of clips for each action is green [24].

3.3.2 The UCF50

A collection of videos used for action recognition makes up the UCF50 dataset. It has fifty distinct action subcategories and YouTube-sourced real-world footage. This dataset is an addition to the eleven categories of different action types included in the YouTube Action dataset (UCF11). UCF50's main goal is to provide the computer vision community a dataset made out of real YouTube videos, as seen in figure 10[25], in contrast to many other action recognition datasets that comprise staged performances by actors.

This dataset will be made available to the general public. The enormous diversity in camera movements, item appearance, posture, object size, perspective, cluttered backdrops, lighting conditions, and other characteristics found in this dataset is the major source of difficulties. A total of 50 groups are provided to cover all of the categories that are accessible, and each of the 25 groups comprises more than four different video clips [25].

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

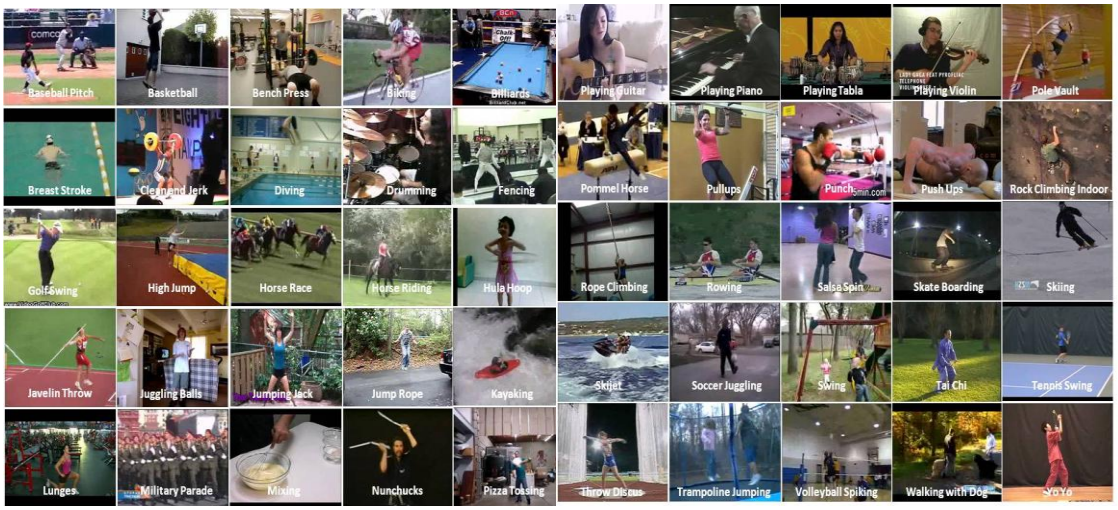


Figure 9. images of the many action types from movies in the UCF50 dataset [25].

3.3.3 Dataset UCF Sports

UCF Sports is a collection of different sporting events collected from broadcast television networks like ESPN and BBC. The ten activities in the dataset include walking, swinging side, swinging bench, lifting, riding a horse, diving, golf swinging, kicking, running, and skateboarding. All 10 activities are shown in one example frame in Figure 10. The dataset is accessible to the public with a human bounding box [26].

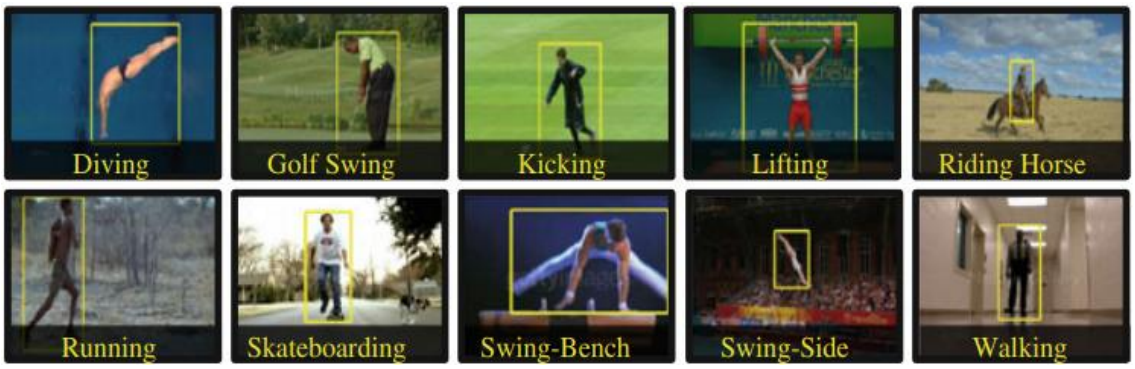


Figure 10. UCF Sport human action [6].

The dataset contains 150 video clips recorded at a fixed frame rate of 10 frames per second with a resolution of 720 x 480 pixels. Table 1 provides a summary of the dataset's properties. Figure 11 illustrates the distribution of the number of clips per action, given that the number of clips in each class varies.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

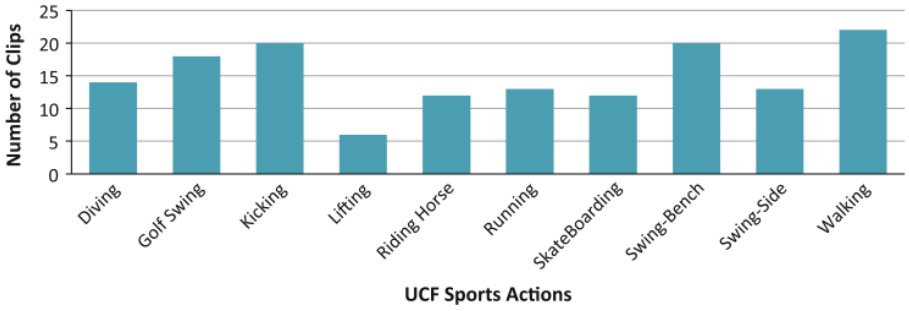


Figure11. The total number of videos for each action class

Table 1. Characteristics of UCF Sports in brief

Actions	10	Total duration	958 s
Clips	150	Frame rate	10 fps
Mean clip length	6.39 s	Resolution	720 × 480
Min clip length	2.20 s	Max num. of clips per class	22
Max clip length	14.40 s	Min num. of clips per class	6

The total duration of blue clips and the average length of the green clips for each action type are shown in figure 12. Certain activities, such as kicking, are clearly shorter than others, such as walking or running, which are significantly longer and more periodic. However, it is clear from the graph that action clips' average lengths across various classes exhibit a lot of overlap. Therefore, evaluating just the time of a single clip would not enough to identify the activity. The following diagram illustrates the division of the total number of clips by activity. Because there are varying amounts of clips in each category [6].

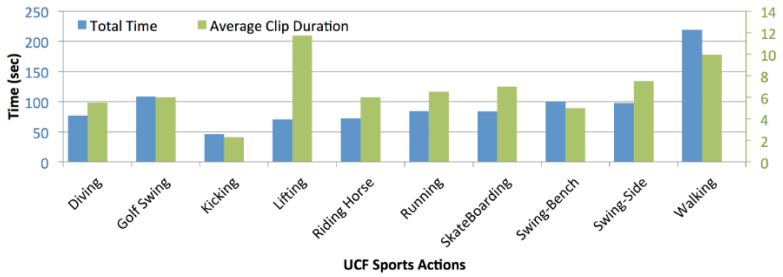


Figure12. Action class video clip lengths are given in blue. Green indicates average action clip length [6].

The UCF Sports Action dataset is a widely used public dataset for action recognition in sports. However, to improve the dataset's accuracy, it is essential to expand it by adding more videos and increasing its diversity. This can be achieved by introducing new sports categories, collecting additional videos for each sport, including footage from various camera angles, enhancing annotation quality, and incorporating videos of athletes with varying skill levels. By implementing these strategies, researchers and developers can improve the dataset's capability and accuracy for various applications, such as action recognition in fitness, and human motion analysis.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

The UCF Sports Action dataset currently contains 10 sports categories The dataset can be extended by adding more sports categories, such as bodyweight squats, handstand push ups, jump rope, lunges, pullups, rope climbing, and push ups. This paper presents an expanded University of Central Florida Sports (EUCF Sport) dataset, comprising 432 video clips gathered from various fitness centres, with a resolution of 720x480 pixels. Figure 13 illustrates the distribution of the number of clips per action, given that the number of clips in each class varies.

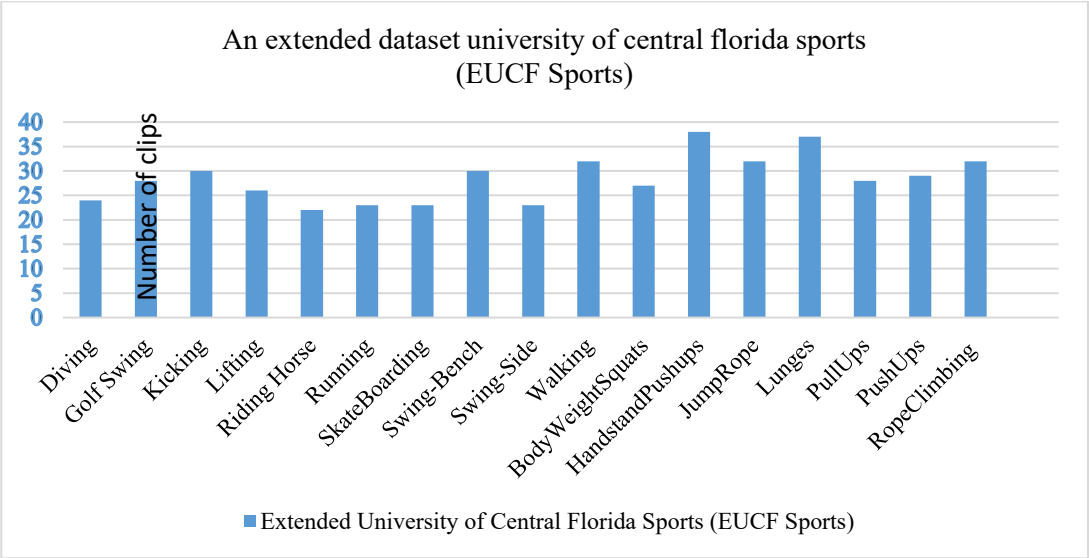


Figure13. The total number of videos for each action class in EUCF Sport

Data pre-processing

In human action recognition, pre-processing the dataset is a crucial step in building an accurate and reliable model. The pre-processing pipeline typically involves several steps such as data cleaning, normalization, feature extraction, data augmentation, data splitting, label encoding, and data visualization. Previously, the model was trained and tested, the quality of the videos needs to be improved. This will help the model better classify and recognise human actions in videos. To reach this goal, different picture processing methods have been built into the suggested model. These techniques include:

Image resizing

Training a deep learning model faster by using smaller image sizes can reduce processing time [27]. However, the RNN architecture requires all input images to be the same standard size, typically ranging from 320x240 pixels to 128x128 pixels in UCF 100. As a result, the EUCF Sport dataset initially includes 720x480 pixel images but then resizes them to 128x128 pixels to match the sizes required by the RNN model.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Colour conversion

Colour conversion from RGB to grayscale is a common pre-processing step in deep learning when working with image data. Converting RGB images to grayscale can help reduce the computational cost of training and inference, as grayscale images have only one channel of pixel values instead of three [28]. The RGB colour model represents the colour of a pixel through three values, namely R, G, and B, each ranging from 0 to 255. Conversely, grayscale images possess a single pixel value that signifies its intensity, ranging from 0 to 1.

Histogram equalization

Histogram equalization is a technique used to adjust the contrast and brightness of an image by redistributing pixel values across a broader range. The goal of histogram equalization is to enhance the visual appearance of an image by increasing the contrast between different pixel intensities. This technique involves altering training images to enhance the training process's dynamism [29].

$$x' = T(x) = \sum_{i=0}^{x} n_i \frac{MAXintensity}{N} \qquad (12)$$

In grayscale images, ni denotes the pixel count with intensity i, while N signifies the total pixel count. The process of histogram equalization converts pixel intensities (x) into new values (x') by fitting the total of a cumulative histogram and a measurement element to fall within the 0-255 intensity range

Data normalization

Data normalization is a common pre-processing step used in image processing to standardize the pixel values of an image. Normalizing the pixel values can help develop the implementation and accuracy of machine learning models trained on the images, as it reduces the impact of differences in brightness and contrast between images. Data normalization between 0 and 1 enhances classification accuracy, significantly affecting feature extraction and classification [31]. Faster neural network training occurs with normalized input images, achieved by centring pixel values around select pre-processing outputs, as illustrated in Figure 7.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

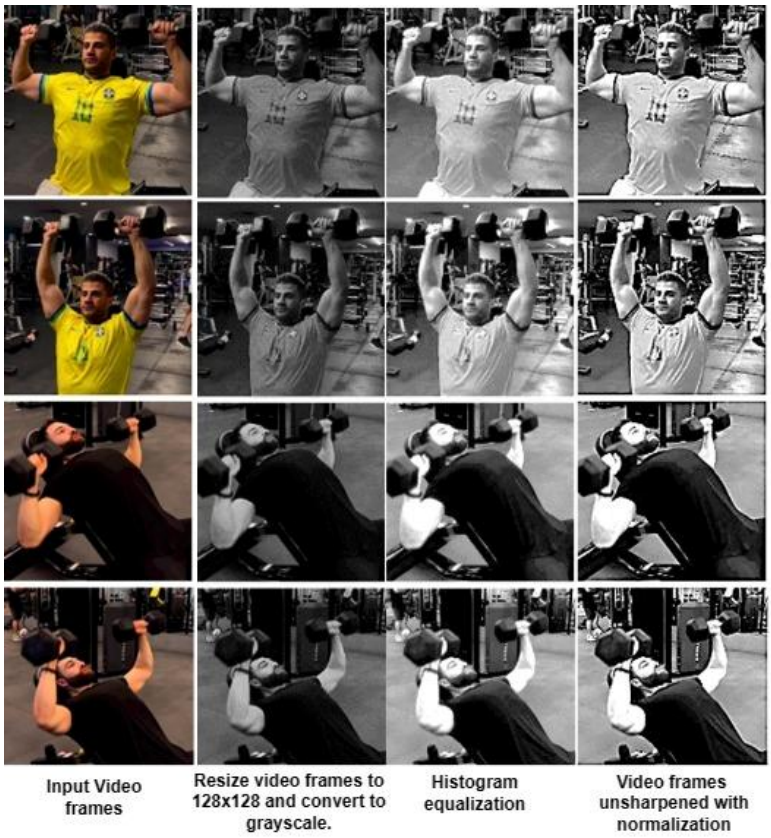


Figure 14. The outcomes of pre-processing algorithms

Data Augmentation

Data augmentation is a technique that generates new data from existing data in order to increase the overall amount of data available. The purpose of this method is to improve the accuracy of machine learning (ML) algorithms by creating multiple varieties of a given dataset [32]. By ensuring that the model does not match the training data too well, this is done during the training phase to avoid over-fitting. To improve training images, images are subjected to horizontal shifts, shearing, rotation, and zooming. Parameters for augmentation are chosen randomly within specific ranges: -10.0, 10.0 % for zooming, 0.0, 0.2 radians counter-clockwise (CCW) and clockwise (CW) for rotation with an angle of 200, and -20.0, 20.0 for horizontal shifts, rotation, zooming, and shearing. Figure 8 provides an example of how augmented images appear.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

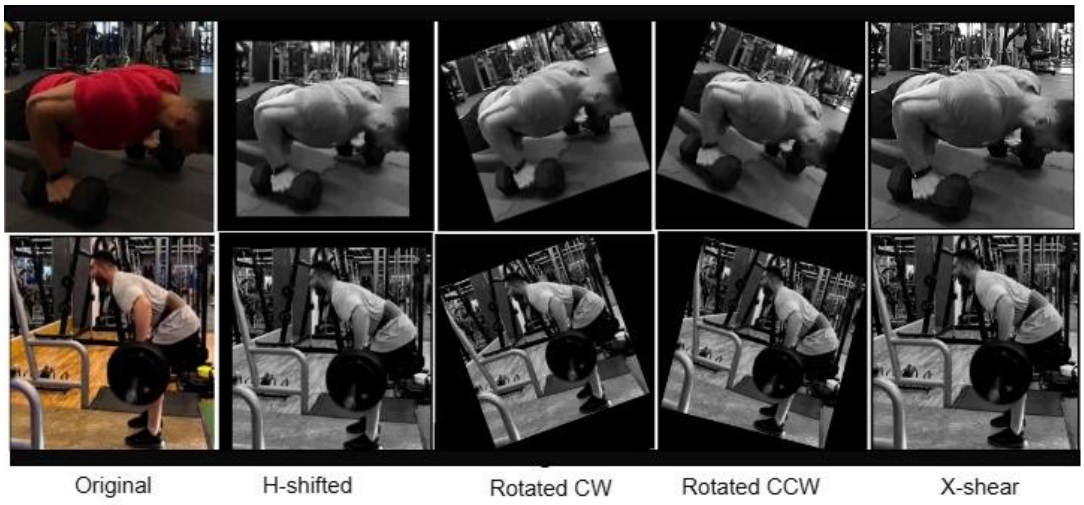


Figure 15. Augmented Human action recognition using EUCF Sports Dataset

Inception V3 for UCF Dataset

Inception-v3 is a deep convolutional neural network (CNN) architecture introduced in 2014 by Google researchers. It was designed to enhance the performance of image classification tasks using a unique architecture that enables better feature extraction. Inception-v3 employs a deep neural network with multiple layers, including convolutional layers, pooling layers, and fully connected layers,. It also utilizes a unique Inception module, which consists of several parallel convolutional layers with different filter sizes and pooling layers combined to extract more diverse and useful features. One of the main advantages of using Inception-v3 for feature extraction is its high accuracy and efficiency [33]. It has demonstrated strong performance in image classification tasks and has been used for various applications, such as object detection, image segmentation, and visual question answering.

To use Inception-v3 for feature extraction, the network can either be trained on a dataset or a pre-trained model that has been trained on a large dataset like ImageNet. Once the network is trained or loaded, the convolutional layers can be used to extract features from input images. These features can then serve as input for another machine learning algorithm or neural network for further processing. There are several different releases: Inception v1 (2014), v2 (2015), v3 (2015), v4 (2016), and Inception-ResNet (2016) [34].

The Inception-V3 convolutional neural network is used for image analysis and object detection. This is the third edition of Google's Inception Convolutional Neural Network. With Inceptionv3, deeper networks can be achieved with fewer parameters, while also keeping the number of parameters under 25 million compared to AlexNet's 60 million [35]. Typically, the Inception module offers three different convolution sizes, as well as a maximum pooling size. After the convolution operation is performed on the previous layer's output, the channels are aggregated, and then nonlinear fusion occurs. This increases the adaptability of the network to different scales and prevents overfitting.

Inception v3 divides large volume integrals into smaller ones using a convolution kernel splitting method. Unlike its predecessors, Inception v1 and v2. Kera's developed Inception v3 as a pre-trained ImageNet network structure. The default image input size is 299x299 with three channels. To divide

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

large volume integrals into smaller convolutions, Inception v3 employs a convolution kernel splitting method, differing from Inceptions v1 and v2. For example, a 3x3 convolution is split into 3x1 and 1x3 convolutions. This splitting method reduces the number of parameters, resulting in faster network training speeds and improved spatial feature extraction [33]. Overall, Inception-v3 is a powerful and efficient architecture for feature extraction and has been widely adopted in the computer vision community for diverse applications.

Evaluation of experimental results

The UCF datasets for action recognition models are used in this part to show experimental assessment of the suggested strategy. Each experiment employs a unique set of optimisations for the LSTM model's four LSTM layers, each of which contains a distinct set of neurons. The SoftMax activation function, a dropout rate of 0.5, and a dense layer in the final layer. Additionally, the GRU model has two layers of GRU, each with a certain number of neurons. Different optimisations were applied for each test, including a SoftMax dense layer activated at the conclusion of the sequence and a dropout rate of 0.5. Two sets comprise the dataset: the training and the validation dataset.80% of the dataset is composed of training data, while the remaining 20% is made up of validation and test data. The model is constructed using the training data throughout the training, validation, and testing processes, and its performance is monitored using the validation data for evaluating its execution during training and testing phases. The results of the models on the UCF101, UCF50, UCF-Sport, and EUCF-Sport datasets correspondingly are presented in table 2, 3, 4, and 5 for RNN models absence of any pre-processing, augmentation, or optimisation. These tables are for RNN models. In addition, Figure 14a illustrates the accuracy and loss analysis of the modified LSTM model. The UCF datasets were analysed without any pre-processing or enhancement.

Table 2. Accuracy of RNN models on UCF101 without pre-process and data augmentation

No. of epochs	Batch size	LSTM model	GRU model
10	16	0.7511756486015319	0.7223529748188781
	32	0.7535298422283172	0.7241170689132691
	64	0.7605882743007659	0.7305796243004532
20	16	0.7605688243007659	0.7578685433007658
	32	0.7741417653369903	0.7703377811369902
	64	0.7936467063732147	0.7815466063732140
30	16	0.8010005033897692	0.8076543151920929
	32	0.818753717102140	0.8097647171402050
	64	0.8384566537826536	0.8158235378265383
40	16	0.8433454653369672	0.8124117653369903
	32	0.8688541768913255	0.8186411768913269
	64	0.8688794740943901	0.8301764740943900
50	16	0.8666572358551133	0.8550882358551025
	32	0.8766774767943456	0.8571176474094390
	64	0.8891374475094377	0.8691176474094390

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 3. Accuracy of RNN models on UCF50 without pre-processing and data augmentation

No. of epochs	Batch size	LSTM model	UCF50 GRU model
10	16	0.773885399341583	0.751430630683898
	32	0.780194699764251	0.779181823730468
	64	0.802819414138794	0.792385643720626
20	16	0.8124682459831237	0.807985076904296
	32	0.8176070048809051	0.806392421960830
	64	0.820330394268035	0.813645824670791
30	16	0.829194699764251	0.813318941593170
	32	0.852281941413879	0.823211743831634
	64	0.864682459831237	0.844885680675506
40	16	0.874638072252273	0.868709397315979
	32	0.880854112148284	0.869400714159011
	64	0.905486562252044	0.877272200584411
50	16	0.904061641931533	0.884198842048645
	32	0.9278895006179809	0.8895834667682647
	64	0.9322526954412460	0.912051653385162

Table 4. Accuracy of RNN models on UCF *Sports* without pre-processing and data

Epochs No	Batch size	LSTM model	GRU model
10	16	0.8008853945469732	0.7724117653369903
	32	0.8091946997642511	0.7873529481887818
	64	0.8228194141369297	0.7988235378653369
20	16	0.8346824535780983	0.8041176533235378
	32	0.8400701237809051	0.8323529601097107
	64	0.8583303942658035	0.8409710832645416
30	16	0.8601946387999303	0.8538235378263775
	32	0.8756872850119414	0.8576470637321472
	64	0.8846843807245983	0.8709882494141599
40	16	0.8894380722598522	0.8780765993369903
	32	0.9068504441121482	0.8985294222831726
	64	0.9154864415622526	0.9046562256754163
50	16	0.9250614411931533	0.91 2352948656225
	32	0.9391895544109807	0.920 0588656225007
	64	0.9561252264569061	0.9260588326845415

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Epochs No	Batch size	LSTM model	GRU model sport
10	16	0.7878853945469732	0.7574117653369903
	32	0.7901946997642511	0.7723529481887818
	64	0.8028194141369297	0.7888235378653369
20	16	0.8246824535780983	0.7941176533235378
	32	0.8380701237809051	0.8023529601097107
	64	0.8703303942658035	0.8309710832645416
30	16	0.8501946387999303	0.8488235378263775
	32	0.8556872850119414	0.8576470637321472
	64	0.8746843807245983	0.8705882494141599
40	16	0.8794380722598522	0.8780765993369903
	32	0.8968504441121482	0.8835294222831726
	64	0.9054864415622526	0.8906562256754163
50	16	0.9150614411931533	0.90 2352948656225
	32	0.9291895544109807	0.910 0588656225007
	64	0.9412522645690617	0.9170588326845415

Table 5. Accuracy of RNN models on EUCF *Sports* without pre-processing and data augmentation

On the UCF101, UCF50, UCF Sports, and EUCF Sports datasets, the LSTM and GRU models showed increased performance after adopting data augmentation and pre-processing approaches. As a consequence, both the training data and the testing data had greater accuracy. Figure 14b demonstrates the accuracy and loss of the updated LSTM model after data pre-processing and enhancement. The results in Tables 6, 7, 8, and 9 show that the 64 and 50 epoch batch sizes provide the best accuracy.

Table 6. The RNN models' accuracy increased when pre-processing and data augmentation methods were used on the UCF101 dataset.

Epochs No	Batch size	LSTM model	GRU model
10	16	0.751176486015456	0.72 41176533699036
	32	0.758529422283223	0.7323529481887817
	64	0.770588243007732	0.74888235378265381
20	16	0.799588243007536	0.75 41176533699036
	32	0.804117653369104	0.77 23529601097107
	64	0.807647063732538	0.79 70588326454163
30	16	0.821000011920990	0.80 88235378265381
	32	0.830764717105090	0.81 76470637321472
	64	0.845823537826678	0.83 05882430076599
40	16	0.8524117653378903	0.84 41176533699036
	32	0.8606411768915369	0.8535294222831726
	64	0.8681764740943680	0.8600588326454163
50	16	0.8770882358551075	0.86 23529481887817
	32	0.8871176474094397	0.87 05882430076599
	64	*0.9161764740943887*	*0.9067058832645416 3*

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 7. The RNN models' accuracy increased when pre-processing and data augmentation methods were used on the UCF50 dataset.

Epochs No	Batch size	LSTM model	GRU model
10	16	0.787885399341583	0.76176471710205
	32	0.790194699764251	0.76911765336990
	64	0.802819414138794	0.77294117927551
20	16	0.824682459831237	0.78964706373214
	32	0.838070048809051	0.81941176891326
	64	0.840330394268035	0.82764706373214
30	16	0.850194699764251	0.83764706373214
	32	0.857281941413879	0.84000001192092
	64	0.874682459831237	0.85895294818878
40	16	0.879438072252273	0.86088235378265
	32	0.896854112148284	0.86058824300765
	64	0.905486562252044	0.88031176891326
50	16	0.925061641931533	0.87941176891326
	32	0.9291895006179809	0.90647063732147
	64	0.9512526954412460	0.92718470637320

Table 8. The RNN models' accuracy increased when pre-processing and data augmentation methods were used on the UCF Sport dataset.

Epochs No	Batch size	LSTM model	GRU model
10	16	0.8395117643198454	0.8214575635654676
	32	0.8486015652908325	0.8241117456824567
	64	0.8599411852359772	0.8312394113567832
20	16	0. 8697722297401429	0.8476464706392755
	32	0. 8747967864706143	0.8573214176891326
	64	0.8862843479156494	0.8687397557790543
30	16	0.8985618347343839	0.8776238755784444
	32	0.9070403984598349	0.8838738326532837
	64	0.9159445083948534	0.8983383209230982
40	16	0.9298438943439859	0.9005684639392002
	32	0.9237745439439349	0.90884674u4932840
	64	0.9454487343443998	0.9189844889803933
50	16	0.9490134665437343	0.9288476338392245
	32	0.9538836523840233	0.9394355720459340
	64	0.9674853489503488	0.9473287328402345

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 9. The RNN models' accuracy increased when pre-processing and data augmentation methods were used on the EUCF Sport dataset.

Epochs No	Batch size	LSTM model	GRU model
10	16	0.8575117643198454	0.8214575635654676
	32	0.8676015652908325	0.8341117456824567
	64	0.8799411852359772	0.8412394113567832
20	16	0. 8867722297401429	0.8566464706392755
	32	0. 8947967864706143	0.8703214176891326
	64	0.9086284347915649	0.8767397557790543
30	16	0.9105618347343839	0.8856238755784444
	32	0.9187040398459834	0.8918738326532837
	64	0.9259445083948534	0.9023383209230982
40	16	0.938438943439859	0.9105684639392002
	32	0.9377745439439349	0.91884674u4932840
	64	0.9504487343443998	0.9369844889803933
50	16	0.9580134665437343	0.9438476338392245
	32	0.9658836523840233	0.9534355720459340
	64	0.9704853489503488	0.9613287328402345

The results showed that the LSTM model was successfully optimised by combining Adam with a SoftMax activation function. This method, when compared to the GRU model, produced results with greater accuracy, depending on the learning rate. The learning rate is used by an optimisation algorithm to choose the step size for all element when it transfers closer to a lowest loss method. Image 14c illustrates the accuracy and loss of the modified LSTM model generated using the optimizer technique. Tables 10, 11, 12, and 13 show that the Adam optimizer had the greatest accuracy with after 50 iterations, the LSTM has only learned at a rate of 0.001. Two training sessions are used in the experiment, each using a different optimisation technique. The number of training epochs is 10, 20, 30, 40, and 50, and there are between 10 and 50 optimisation trials. There are three different learning rates used: 0.001, 0.01, and 0.1. Using an LSTM rate of learning (0.001) within 50 epochs yields the top level.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 10. The accuracy of the RNN model while using the Adam optimizer method on UCF101

Epochs No	Learning rate	LSTM model	GRU model
10	0.001	0.799411792755127	0.7611764860153198
	0.01	0.7629411852359772	0.7441176652908325
	0.1	0.7570588266849518	0.7229411852359772
20	0.001	0.8370588445663452	0.8297522974014282
	0.01	0.8170588445663452	0.8064706134796143
	0.1	0.8109412031173706	0.782843479156494
30	0.001	0.8805882549285889	08685612144470215
	0.01	0.8623529601097107	0.8404058963775635
	0.1	0.8535294163227081	0.8251176533699036
40	0.001	0.9114706015586853	0.8964706134796143
	0.01	0.9058823704719543	0.8877743158340454
	0.1	0.8982353007793426	0.8841764979362488
50	0.001	0.9519412031173706	0.92901477575302124
	0.01	0.9389411911964417	0.90988235378265381
	0.1	0.9195823823928833	0.90667280974388123

Table 11. The accuracy of the RNN model of the Adam optimizer technique on UCF50

Epochs No	Learning rate	LSTM model	GRU model
10	0.001	0.8105882549285889	0.7611764852908325
	0.01	0.8023529601097107	0.7441176660153198
	0.1	0.7955294163227081	0.7229411724014282
20	0.001	0.8614706015586853	0.8297522978523597
	0.01	0.8458823704719543	0.8034796143156499
	0.1	0.8382353007793426	0.7828434799464706
30	0.001	0.9019412031173706	086856121444702151
	0.01	0.8941191196884417	0.8404017653369906
	0.1	0.8895823823928833	0.8251589637756353
40	0.001	0.9341176660153198	0.8964706134796143
	0.01	0.9192294117240142	0.8877743404542488
	0.1	0.9117522978523597	0.8841764979362124
50	0.001	0.9634796143156499	0.9490147757530353
	0.01	0.9502843479946470	0.9298823533813467
	0.1	092785612144470210	0.9076265097438812

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 12. The performance of the RNN model on the UCF Sports data using the Adam optimizer approach

Epochs No	Learning rate	LSTM model	GRU model
10	0.001	0.9004509340535398	0.877348782358209
	0.01	0.8946458043952365	0.8708874875304983
	0.1	0.8890503459948394	0.8632307034598023
20	0.001	0.9224843348784485	0.9008740459567857
	0.01	0.9198723874734087	0.8880582502830223
	0.1	0.9082343239898332	0.8823423233656445
30	0.001	0.9448943873473470	0.9272432493232400
	0.01	0.9398438734875435	0.9132389873272322
	0.1	0.9288384387747875	0.908398834879328
40	0.001	0.9583487347834473	0.9447769832080983
	0.01	0.9528743873487534	0.9392349827349828
	0.1	0.9493934873457834	0.9348530485087348
50	0.001	0.9808343478934838	0.9693458740934900
	0.01	0.9764987347345873	0.9556678554424456
	0.1	09649834873483483	0.9494546867545343

Table 13. The performance of the RNN model on the EUCF Sports data using the Adam optimizer approach

Epochs No	Learning rate	LSTM model	GRU model
10	0.001	0.9123509340535398	0.8897348782358209
	0.01	0.9046458043952365	0.8808874875304983
	0.1	0.8908503459948394	0.8754547764598023
20	0.001	0.9374843348784485	0.9108740459567857
	0.01	0.9288723874734087	0.9020582502830223
	0.1	0.9202343239898332	0.8967423233656445
30	0.001	0.9588943873473470	0.9299932493232400
	0.01	0.9508438734875435	0.9232389873272322
	0.1	0.947384387747875	0.9198398834879328
40	0.001	0.9783487347834473	0.9537769832080983
	0.01	0.9648743873487534	0.9442349827349828
	0.1	0.9593934873457834	0.9388530485087348
50	0.001	0.9978343478934838	0.9853458740934900
	0.01	0.9864987347345873	0.9696678554424456
	0.1	09809834873483483	0.9604546867545343

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

In order to train a deep learning model, it is necessary to make modifications to the weights of each epoch and discover the value of optimal for the loss function. The neural network optimizer makes changes to the parameters of the network, including the weights and the rate of learning. This makes the system more accurate and cuts down on losses around the world. The most accurate model is one that combines an LSTM with GSD.

Nevertheless, there are several factors that can impact this outcome, such as the type of data utilised, the results of pre-processing, the architecture chosen, and the tuning parameters applied. In Table 14, Table 15, Table 16 and Table 17, the performance of the proposed models on the UCF101, UCF50, UCF Sports and EUCF Sports datasets, respectively, is presented for RNN models, after implementing video pre-processing, data augmentation, and optimisation techniques.

Table 14. The accuracy of the assessment models used for the SGD optimizer approach on UCF101

Epochs No	Learning rate	LSTM model	GRU model
10	***0.001***	*0.820411792755127*	*0.8011764860153198*
	0.01	*0.8129411852359772*	*0.7841176652908325*
	0.1	*0.8070588266849518*	*0.7729411852359772*
20	***0.001***	*0.8470588445663452*	*0.8297522974014282*
	0.01	*0.8270588445663452*	*0.8064706134796143*
	0.1	*0.8109412031173706*	*0.8002843479156494*
30	***0.001***	*0.8805882549285889*	*0.8605612144470215*
	0.01	*0.8623529601097107*	*0.8504058963775635*
	0.1	*0.85352941632270813*	*0.8451176533699036*
40	***0.001***	*0.9014706015586853*	*0.8764706134796143*
	0.01	*0.8988823704719543*	*0.8677743158340454*
	0.1	*0.89223530077934265*	*0.8681764979362488*
50	***0.001***	*0.9299412031173706*	*0.91901477575302124*
	0.01	*0.9189411911964417*	*0.90388235378265381*
	0.1	*0.9095823823928833*	*0.90067280974388123*

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 15. The accuracy of the SGD optimizer technique's assessment models on UCF50

Epochs No	Learning rate	LSTM model	GRU model
10	***0.001***	*0.835882549285889*	*0.8025612144470215*
	0.01	*0.81235296015097107*	*0.7954058963775635*
	0.1	*0.80352941632270813*	*0.7801176533699036*
20	***0.001***	*0.8614706015586853*	*0.8364706134796143*
	0.01	*0.8459411852359772*	*0.8241176652908325*
	0.1	*0.8470588266849518*	*0.8189411852359772*
30	***0.001***	*0.8970588445663452*	*0.8697522974014282*
	0.01	*0.8870588445663452*	*0.8564706134796143*
	0.1	*0.8709412031173706*	*0.8499843479156494*
40	***0.001***	*0.9203529601097107*	*0.894058963775635*
	0.01	*0.90352941632270813*	*0.8851176533699036*
	0.1	*0.9014706015586853*	*0.8704706134796143*
50	***0.001***	*0.9488823704719543*	*0.9177743158340454*
	0.01	*0.93223530077934265*	*0.9091764979362488*
	0.1	*0.92705884455663452*	*0.9004706134796143*

Table 16. The accuracy of the SGD optimizer technique's assessment models on UCF Sports

Epochs No	Learning rate	LSTM model	GRU model
10	***0.001***	*0.875882549285889*	*0.8605612144470215*
	0.01	*0.86235296015097107*	*0.8504058963775635*
	0.1	*0.84352941632270813*	*0.8451176533699036*
20	***0.001***	*0.9014706015586853*	*0.8864706134796143*
	0.01	*0.8959411852359772*	*0.8741176652908325*
	0.1	*0.8870588266849518*	*0.8689411852359772*
30	***0.001***	*0.9210588445663452*	*0.9097522974014282*
	0.01	*0.9170588445663452*	*0.8964706134796143*
	0.1	*0.91109412031173706*	*0.8919843479156494*
40	***0.001***	*0.9503529601097107*	*0.9384058963775635*
	0.01	*0.93352941632270813*	*0.9251176533699036*
	0.1	*0.9284706015586853*	*0.9104706134796143*
50	***0.001***	*0.9828823704719543*	*0.9677743158340454*
	0.01	*0.97223530077934265*	*0.9481764979362488*
	0.1	*0.95705884455663452*	*0.9404706134796143*

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 17. The accuracy of the SGD optimizer technique's assessment models on EUCF Sports

Epochs No	Learning rate	LSTM model	GRU model
10	***0.001***	*0.885882549285889*	*0.8785612144470215*
	0.01	*0.8723529601509718*	*0.8704058963775635*
	0.1	*0.86352941632270813*	*0.8651176533699036*
20	***0.001***	*0.9104706015586853*	*0.8964706134796143*
	0.01	*0.9059411852359772*	*0.8891176652908325*
	0.1	*0.8990588266849518*	*0.8849411852359772*
30	***0.001***	*0.9380588445663452*	*0.9197522974014282*
	0.01	*0.9270588445663452*	*0. 9084706134796143*
	0.1	*0.91809412031173706*	*0.9019843479156494*
40	***0.001***	*0.9603529601097107*	*0.9484058963775635*
	0.01	*0.95952941632270813*	*0.9331176533699036*
	0.1	*0.94114706015586853*	*0.9204706134796143*
50	***0.001***	*0.9898823704719543*	*0.9707743158340454*
	0.01	*0.98123530077934265*	*0.9641764979362488*
	0.1	*0.97205884455663452*	*0.9564706134796143*

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

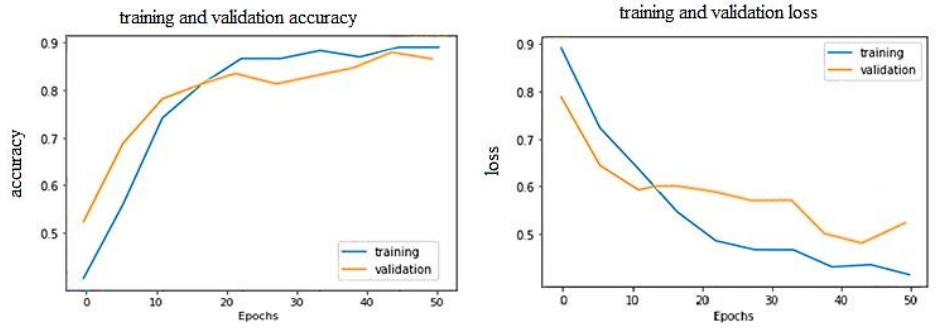


Figure 14a: Loss and accuracy of LSTM models when no preprocessing and no augmentation of dataset.

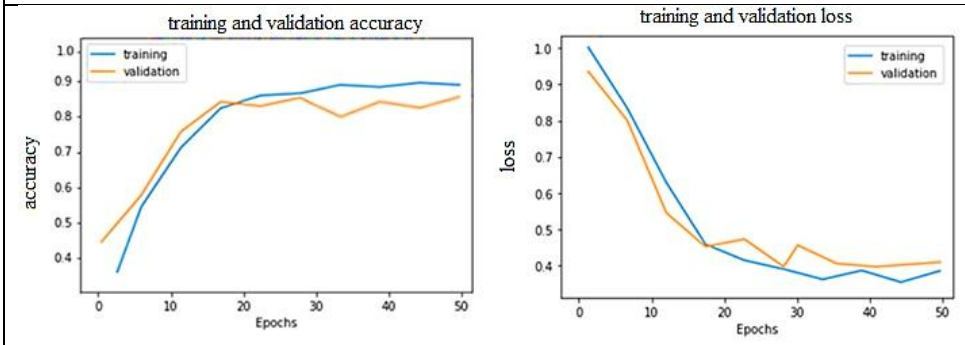


Figure 14b: The effect of pre-processing and augmentation on the accuracy and loss of LSTM models

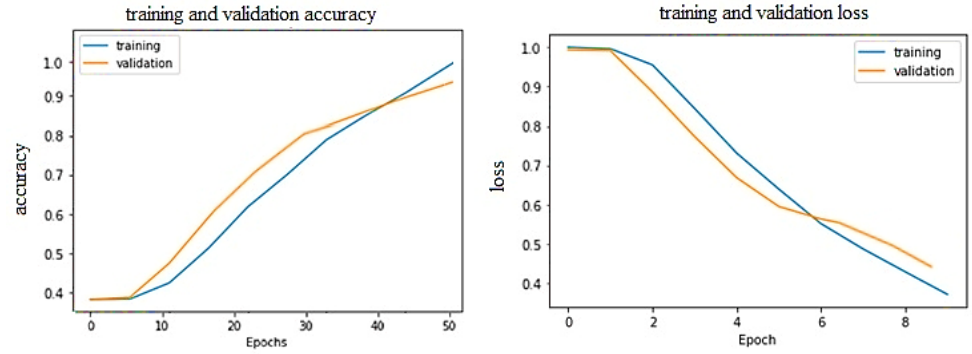


Figure 14c: The result of preprocessing, optimization and data augmentation on the accuracy and loss of LSTM model.

Figure 14: LSTM model accuracy and loss on EUCF Sports dataset

In the past, there have been attempts to classify videos, especially when it came to human activity movies. The UCF Sports and EUCF Sports files have gotten a lot of attention because they are bigger and can be used by anyone. When looking at these datasets, researchers have often chosen the RNN design, which is known for its accuracy. In Table 18, you can see how the suggested models compare to recent studies in terms of how accurate they are. It is clear that the suggested methods, which use LSTM with Adam optimisation, got an average accuracy of 99% on UCF Sports, while GRU got 98%, which outperforms other deep learning techniques.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 18. The accuracy of the proposed model is compared to that of relevant previous studies.

No	Paper	Year	Method	Class of Architecture	Dataset	Accuracy
1	[36]	2015	static and motion features	CNN	UCF Sport	91.9
2	[37]	2016	Spatio-temporal	LSTM-CNN	UCF-101	86.9
3	[38]	2016	LRCN		UCF-101	82.34
4	[39]	2017	SP-CNN	CNN	UCF-101	91.6
5	[40]	2017	CNNs and LSTMs	CNN-LSTMs	UCF Sport UCF-101	92.20 89.20
6	[41]	2018	QST-CNN-LSTM	CNN-RNN	UCF Sport UCF-101	93.20 89.70
7	[42]	2019	CNN+LSTM	CNN-LSTM	UCF Sports	85%.
8	[43]	2020	GMM + KF + GRNN	RNN	UCF Sport UCF 101	89.01 89.30
9	[44]	2020	Semi-CNN Semi-CNN Semi-CNN	ResNet VGG-16 DenseNet	UCF-101	89.00 82.58 77.72
10	[45]	2021	GMM+PKF+GRNN	RNN	UCF Sport UCF-101	89.98% 90.31%
11	**Proposed Models**		**(LSTM)** Our method ($\lambda = 0.01$)	RNN	EUCF Sport UCF Sport UCF 50 UCF-101	**99.78** 98.08 96.34 95.19
			(GRU) Our method ($\lambda = 0.01$)		EUCF Sport UCF Sport UCF 50 UCF-101	**98.53** 96.93 94.90 92.90

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

The Scientific Journal of Cihan University – Sulaimaniya **PP: 82-117**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Conclusions

This study looks at how accurate deep-learning techniques like modified LSTM and GRU models can be used to classify ways to recognize human actions. The LSTM model was modified to have four hidden layers, while the GRU model had two hidden layers. Additionally, pre-processing methods were used to minimise overfitting by applying modifications including translation, scaling, and rotation to the UCF101, UCF50, UCF Sports, and EUCF Sports datasets. The results demonstrated the effectiveness of optimising an LSTM model using the Adam optimizer and a SoftMax activation function. After 50 iterations on UCF101, the LSTM model achieved a maximum accuracy of 95.19% using the Adam optimizer with a learning rate of 0.001. The GRU model, on the other hand, reached a peak accuracy of 92.90% after 50 epochs and 0.001 learning rate. On the UCF50 dataset, the LSTM model with the Adam optimizer was the most accurate, with a 96.34% accuracy rate after 50 epochs and a 0.001 learning rate. At 50 epochs, the GRU model learned at a rate of 0.001 and got as close as 94.90% of the time. Also, the model with a learning rate of 0.001 reached a peak accuracy of 96.93% after 50 epochs, while the LSTM model with the Adam optimizer with the same learning rate reached a peak accuracy of 98.80%. Lastly, on the EUCF Sports dataset, the most accurate model was the LSTM model with the Adam optimizer and a learning rate of 0.001. After 50 epochs, it was right 99.78% of the time. On the other hand, after 50 epochs and a learning rate of 0.001, the GRU model was most accurate to the tune of 98.53%.

Conflicts of Interest

The authors state that there are no conflicts of interest.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

References

[1] I. Jegham, A. Ben, I. Alouani, and M. Ali, "Forensic Science International : Digital Investigation Vision-based human action recognition : An overview and real world challenges," Forensic Sci. Int. Digit. Investig., vol. 32, p. 200901, 2020, doi: 10.1016/j.fsidi.2019.200901.

[2] Angelov, P.P. et al. (2016) "Autonomous data density based Clustering Method," 2016 International Joint Conference on Neural Networks (IJCNN) [Preprint]. Available at: https://doi.org/10.1109/ijcnn.2016.7727498.

[3] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," Comput. Intell. Neurosci., vol. 2018, 2018, doi: 10.1155/2018/7068349.

[4] N. Ahmed, K. Nouduri, and K. Palaniappan, "A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network," no. October, 2020, doi: 10.1109/AIPR50011.2020.9425332.

[5] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," Appl. Sci., vol. 7, no. 1, 2017, doi: 10.3390/app7010110.

[6] Muhamad, Azhee W., & Mohammed, Aree A. (2022). "Review on recent Computer Vision Methods for Human Action Recognition". Advances in Distributed Computing and Artificial Intelligence Journal, pp. 361- 379,2021, DOI: https://doi.org/ 10.14201/ ADCAIJ2021104361379.

[7] M. M. Hossain Shuvo, N. Ahmed, K. Nouduri, and K. Palaniappan, "A hybrid approach for human activity recognition with support vector machine and 1d convolutional neural network," Proc. - Appl. Imag. Pattern Recognit. Work., vol. 2020-October, no. October, 2020, doi: 10.1109/AIPR50011.2020.9425332.

[8] N. ur R. Malik, S. A. R. Abu-Bakar, U. U. Sheikh, A. Channa, and N. Popescu, "Cascading Pose Features with CNN-LSTM for Multiview Human Action Recognition," Signals, vol. 4, no. 1, pp. 40–55, 2023, doi: 10.3390/signals4010002.

[9] J. Zhang, L. Zi, Y. Hou, M. Wang, W. Jiang, and D. Deng, "A Deep Learning-Based Approach to Enable Action Recognition for Construction Equipment," Adv. Civ. Eng., vol. 2020, 2020, doi: 10.1155/2020/8812928.

[10] I. Sipiran and B. Bustos, "Harris 3D: A robust extension of the Harris operator for interest point detection on 3D meshes," Vis. Comput., vol. 27, no. 11, pp. 963–976, 2011, doi: 10.1007/s00371-011-0610-y.

[11] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," Pattern Recognit. Lett., vol. 131, pp. 293–299, 2020, doi: 10.1016/j.patrec.2020.01.010.

[12] Kurnianggoro, L., Wahyono and Jo, K.-H. (2018) "A survey of 2D shape representation: Methods, evaluations, and future research directions," Neurocomputing, 300, pp. 1–16. Available at: https://doi.org/10.1016/j.neucom.2018.02.093.

[13] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," Multimed. Tools Appl., vol. 79, no. 41–42, pp. 30509–30555, 2020, doi: 10.1007/s11042-020-09004-3.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

[14] Naeem, H.B. et al. (2020) "Multiple batches of motion history images (MB-mhis) for multi-view human action recognition," Arabian Journal for Science and Engineering, 45(8), pp. 6109–6124. Available at: https://doi.org/10.1007/s13369-020-04481-y.

[15] L. Cai, C. Liu, R. Yuan, and H. Ding, "Human action recognition using Lie Group features and convolutional neural networks," Nonlinear Dyn., vol. 99, no. 4, pp. 3253–3263, 2020, doi: 10.1007/s11071-020-05468-y.

[16] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," Appl. Sci., vol. 7, no. 1, 2017, doi: 10.3390/app7010110.

[17] Li, C. et al. (2019) "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," Pattern Recognition, 85, pp. 161–171. Available at: https://doi.org/10.1016/j.patcog.2018.08.005.

[18] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., &Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE access, 6, 1155-1166.

[19] Chu, W. et al. (2019) "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," IEEE Transactions on Multimedia, 21(1), pp. 246–255. Available at: https://doi.org/10.1109/tmm.2018.2846411.

[20] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Networks Learn. Syst., vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/ TNNLS.2016.2582924.

[21] Ida, Y. and Fujiwara, Y. (2020) "Improving generalization performance of adaptive learning rate by switching from block diagonal matrix preconditioning to SGD," 2020 International Joint Conference on Neural Networks (IJCNN) [Preprint]. Available at: https://doi.org/10.1109/ijcnn48605.2020.9207425.

[22] Wang, Z. et al. (2022) 'Human action recognition based on improved two-stream convolution network', Applied Sciences, 12(12), p. 5784. doi:10.3390/app12125784.

[23] Wang, X. et al. (2019) 'I3d-LSTM: A new model for human action recognition', IOP Conference Series: Materials Science and Engineering, 569(3), p. 032035. doi:10.1088/1757-899x/569/3/032035.

[24] Muhamad, A.W. and Mohammed, A.A. (2023) 'A comparative study using improved LSTM /GRU for human action recognition', 15th March 2023. Vol.101. No 5, 101(5), pp. 1863–1879. doi:10.21203/rs.3.rs-2380406/v1.

[25] Reddy, K.K. and Shah, M. (2013) "Recognizing 50 human action categories of web videos," Machine Vision and Applications, 24(5), pp. 971–981. Available at: https://doi.org/10.1007/s00138-012-0450-4.

[26] Sultana, F., Sufian, A. and Dutta, P. (2020) "A review of object detection models based on Convolutional Neural Network," Advances in Intelligent Systems and Computing, pp. 1–16. Available at: https://doi.org/10.1007/978-981-15-4288-6_1.

[27] D. S. David and M. Samraj, "Artech Journal of Effective Research in Engineering and Technology (AJERET) ISSN : 2523-6164 A Comprehensive Survey of Emotion Recognition System in Facial Expression," no. 3, pp. 76–81, 2020.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

[28] K.G. Dhal, A. Das, S. Ray, J. Gálvez, and S. Das, "Histogram equalization variants as optimization problems: a review," Archives of Computational Methods in Engineering, vol. 28, no. 3, pp.1471-1496, 2021.

[29] K. Munadi, K. Muchtar, N. Maulina, and B. Pradhan, "Image enhancement for tuberculosis detection using deep learning," IEEE Access, vol. 8, pp.217897-217907, 2020.

[30] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," Inf. Fusion, vol. 53, no. March 2019, pp. 80–87, 2020, doi: 10.1016/j.inffus.2019.06.014.

[31] Qin, Zhen et al. (2020) 'Imaging and fusing time series for wearable sensor-based human activity recognition', Information Fusion, 53, pp. 80–87. doi:10.1016/j.inffus.2019.06.014.

[32] Wang, Y. et al. (2019) "Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network," Applied Soft Computing, 74, pp. 40–50. Available at: https://doi.org/10.1016/j.asoc.2018.10.006.

[33] Alom, M.Z. et al. (2018) "Improved inception-residual convolutional neural network for object recognition," Neural Computing and Applications, 32(1), pp. 279–293. Available at: https://doi.org/10.1007/s00521-018-3627-6.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308

[35] Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. Learning to track for spatio-temporal action localization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3164–3172

[36] Mahasseni, B.; Todorovic, S. Regularizing long short-term memory with 3D human-skeleton sequences for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3054–3062.

[37] Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.

[38] Yu, S. et al. (2016) "Stratified pooling based deep convolutional Neural Networks for Human Action Recognition," Multimedia Tools and Applications, 76(11), pp. 13367–13382. Available at: https://doi.org/10.1007/s11042-016-3768-5.

[39] Gammulle, H. et al. (2017) "Two stream LSTM: A deep fusion framework for Human Action Recognition," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) [Preprint]. Available at: https://doi.org/10.1109/wacv.2017.27.

[40] Meng, B., Liu, X.J. and Wang, X. (2018) "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos," Multimedia Tools and Applications, 77(20), pp. 26901–26918. Available at: https://doi.org/10.1007/s11042-018-5893-9.

[41] M.R., A., Makker, M. and Ashok, A. (2019) "Anomaly detection in surveillance videos," 2019 26th International Conference on High Performance Computing, Data and Analytics Workshop (HiPCW) [Preprint]. Available at: https://doi.org/10.1109/hipcw.2019.00031.

DOI: http://dx.doi.org/10.25098/9.1.30
Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)

[42] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 4, pp. 447–453, 2020, doi: 10.1016/j.jksuci.2019.09.004.

[43] Leong, M.C. et al. (2020) "Semi-cnn architecture for effective spatio-temporal learning in action recognition," Applied Sciences, 10(2), p. 557. Available at: https://doi.org/10.3390/app10020557.

[44] Kumar, B.S., Raju, S.V. and Reddy, H.V. (2021) "Human action recognition using a novel deep learning approach," IOP Conference Series: Materials Science and Engineering, 1042(1), p. 012031. Available at: https://doi.org/10.1088/1757-899x/1042/1/012031.

DOI: http://dx.doi.org/10.25098/9.1.30

Distributed under the terms and conditions of the License 4.0 (CC BY-NC-ND 4.0)