Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

DKJ, XL, SN, and MP either did not respond directly or could not be reached.

Modeling of human action recognition using hyperparameter tuned deep learning model

Deepak Kumar Jain,[a] Xue Liu,[b,*] Subramani Neelakandan,[c] and Mohan Prakash[d]

[a]Chongqing University of Posts and Telecommunications, College of Automation, Chongqing, China
[b]Wuhan Donghu University, Basic Course Department, Wuhan, China
[c]R.M.K. Engineering College, Department of Computer Science and Engineering, Tamil Nadu, India
[d]Vellore Institute of Technology, School of Computer Science and Engineering, Vellore, Tamilnadu, India

Abstract. Recently, human action recognition (HAR) has become an important focus of computer science research because of its applications in surveillance, robotics, sentiment analysis, and other areas. Human activity classification is a time-consuming operation, especially when photos are cluttered, and the background is unclear. In addition, conventional machine learning models fail to achieve robust performance due to the increasing number of activities. Comparatively, deep learning (DL) approaches help to automatically learn and describe the necessary features of the input data with low manual work and robust discriminant abilities. In addition, a preprocessing stage is included in this study's HAR utilizing hyperparameter tuned DL (HAR-HPTDL) model that removes undesired background and improves the quality of the input. The model also implements a bidirectional long short-term memory model as a feature extractor, the sparrow search algorithm to tune the hyperparameters, and a SoftMax layer for the effective classification of human actions. In addition, the curse of dimensionality can be overcome via entropy-based feature reduction and Chi square-based feature selection. Based on a variety of measures, the HAR-HPTDL methodology has been put through its paces with other published techniques. The results indicate that the HAR-HPTDL technique outperforms current state-of-the-art techniques in simulations. The outcome of this work demonstrates that an HAR-HPTDL model may achieve comparable or even superior recognition accuracy of 0.949 than the prior best deep classifier(s) on all databases with proper parameter optimization. [DOI: 10.1117/1.JEI.32.1.011211]

Keywords: deep learning; human activity recognition; classification; hyperparameter optimization; feature reduction.

Paper 220387SS received May 29, 2022; accepted for publication Aug. 8, 2022; published online Sep. 14, 2022.

1 Introduction

Human action recognition (HAR) has sparked increased interest among scientists considering its wide range of real-time applications, such as smart and intelligent surveillance systems, human-to-machine or human-to-object interactions, virtual reality/augmented reality, content-based data retrieval, autonomous driving, games, and health care systems.[1] In recent times, the requirements for HAR and pose estimations have considerably increased to evade intimate contact during the pandemic and offer convenient interactions for rehabilitees.[2] A human action recognition system (HARS) must be constantly updated due to the ever-changing technology in the field and the multidisciplinary nature of HAR. The development of HAR systems from the standpoint of computer vision has a significant connection to computer vision applications. Most CV applications have a strong connection to HAR tasks. The HAR method emphasizes the recognition of an activity precisely regarding a kind of behavior initiated in sequences of frames in the video

*Address all correspondence to Xue Liu, lxue1120@sina.com

1017-9909/2022/$28.00

captured. Human activity recognition (HAR) can help a variety of applications, including health care and smart home applications. Because of the rapid growth of wireless sensor networks, a great amount of data may be collected to recognize human behaviors using various types of sensors. Traditional machine learning (ML) algorithms necessitate the manual extraction of representative features from data. On the other side, manual feature engineering calls for specialized knowledge and is doomed to ignore implicit features. In a number of challenging academic domains, deep learning (DL) has recently achieved extraordinary success. Through the use of DL, it is feasible to automatically uncover representative traits in big datasets. It has the potential to be an excellent tool for monitoring human activity.

The automated detection of human actions via CV has become more efficient in recent years and consequently with the rapidly increasing requirements in different sectors.[3] This technology is pertinent for monitoring activities in smart homes, health care systems, security and environmental regulations, driver assistance systems, and autonomous vehicles to achieve automated recognition of abnormal behaviors, such as terrorist or criminal activity, that must be reported to the appropriate authority. Moreover, services such as home automation, intelligent meeting rooms, and entertainment environment can improve human interactions with computer and personal digital assistant, which is especially significant during the COVID19 out-break when social distancing was required. However, complicated HAR techniques utilizing CVs demand a higher computational cost, as capturing videos are affected by visibility, light, orientation, and scale.[4] Hence, for reducing the computation cost, an HAR method must be able to effectively recognize a subject's activity with minimum data, which is obtained online and measured in real-world. As a result, frame index information may be used, and the body's position might be represented by a series of direction rectangles in human pose estimations. Every frame's state descriptors are created by combining rectangle positions and directions into a histogram background subtraction (BGS), which uses the backdrop as an offset, and approaches such as the motion boundary histogram, the histogram of oriented gradients (HOG) and histogram of optical flow. Skeleton models can be used to characterize human activities by capturing the locations of body parts such as the arms and hands. Various ML approaches have also been introduced for recognizing actions and address the abovementioned problems. Nevertheless, these approaches still suffer from individual deficiencies, weaknesses, and strengths.

The primary objectives in the HAR field include the analysis, representation, and detection of human activities.[5] While ML methods are extensively utilized to attain these objectives, they fail to demonstrate the strength of action recognition and are easily affected by changes in the camera angle. In recent years, advances in technology have enabled the training and use of deep neural network frameworks capable of learning representations from data without the use of handcrafted features/rules, especially when their accuracy improves rapidly as more data are provided. Several open-source datasets have also emerged in the field of HAR, allowing for the testing of new frameworks and activity representation in real-time scenarios.[6] In addition to the above-mentioned problems, the development of new deep frameworks and their applications in real-time situations are critical focuses of HAR studies.

Convolutional neural networks (CNNs) are a type of deep neural network that efficiently categorize objects by combining filtering and layers.[7] With the convolutional technique, each input image undergoes a single time step before being sent to the long short-term memory (LSTM). An image's multiple portions can be mapped using filter maps, the most critical hyperparameter. For instance, the RNN method could be used to solve a few challenges in HAR. Specifically, RNN features a recursive loop that keeps track of the gathered data but only keeps one previous step, which is considered a disadvantage. In comparison, LSTM can store data from multiple phases in a sequential order and avoids the challenge of vanishing and exploding gradients, unlike RNN, for maintaining long-term dependency. This study proposes a new HAR model using hyperparameter-tuned DL (HAR-HPTDL), which implements a preprocessing stage to remove noise via the Weiner filtering (WF) technique. An SSA with an effective feature extractor for HAR was used.

Models of bidirectional long short-term memory (BiLSTM) are employed. In addition, the curse of dimensionality is addressed by combining entropy-based feature reduction with Chi square-based feature selection, while a SoftMax (SM) layer is applied to determine human

actions in an effective way. At last, various simulation analysis and comparative studies with existing techniques were performed to guarantee the supremacy of the HAR-HPTDL technique.

2 Review of Existing HAR Approaches

This section performs a brief survey of the existing HAR approaches presented in the literature. Serpush and Rezaei[6] addressed the challenge of preprocessing with an automatic election of illustrative frame from the input sequence, by extracting the main characteristics of the frames instead of the whole features. They proposed a hierarchical method that utilizes HOG and BGS, following the application of DNN and skeletal modeling. Moreover, the method integrates CNN and LSTM recursive networks for FS to maintain the prior data and SM KNN classifiers to label human actions.

Tasnim et al.[7] proposed an spatial-temporal image formation (STIF) method for a three-dimensional (3D) skeleton joint by taking spatial data and temporal modifications for discrimination actions. They conducted TL (pretrained methods, such as DenseNet121, ResNet18, and MobileNetV2, using ImageNet datasets) for extracting discriminative features and to calculate the presented technique with various fusion methods. They also examined the effects of three fusion models, including maximization, elementwise average, and multiplication, on the efficiency variations of HAR. Jaouedi et al.[8] introduced a new HAR method based on feature extraction and video footage. A motion feature is achieved by human action tracing with the GMM and KF methods, while other features dependent on each visual characteristic in all video sequence frames are extracted with the RNN method using GRU. The major benefits of this new method are the extraction and analysis of each feature at all times and in all video frames.

Yu et al.[9] used ensemble DL to dissect the body position and recognize the background data in photographs to achieve HAR. With the use of the pretrained NCNN method, they created an end-to-end NCNN technique. NCNN can learn separate spatial and channel features using parallel branches, which can enhance the efficiency of the technique. As a result, they presented an end-to-end DELWO approach for maximizing the benefits of nonsequential topology that helps to manually merge deep data collected from a variety of sources. Finally, they devised the DELVS model, which combines a number of deep approaches with weighted coefficients to produce the best possible forecasts.

Sargano et al.[10] proposed a new HAR approach that uses pretrained CNN models as a source framework to extract features from the target dataset, followed by a hybrid SVM and KNN for action categorization. They reported that previously learned CNN-based representation on large-scale annotated datasets is effectively transportable to HAR tasks with constrained training datasets. A CNN-based HAR method for ADLs was described by Mathe et al.[11] DFT images are used to train a neural network, meaning that all HARs are ultimately represented by images. They generated 3D skeleton locations of human joints from raw RGB sequences and improved them with depth data. Further, 3-1D signals were obtained to characterize the mobility of every joint, exhibiting its coefficient in 3D Euclidean spaces.

With the support of Conv-LSTM and FC-LSTM, Zhang et al.[12] handled the HAR problem with unique attentions. Then, the STDAN is generally made up of fusion modules and feature extraction where attention is developed. Mukherjee et al.[13] presented EnsemConvNet, which combines CNN-LSTM, CNN-Net, and Encoded-Net. These three classification methods are based on a simple one-dimensional CNN, but differ in terms of the number of kernel size, dense layers, and the framework's major variance. Each approach accepts time sequence data as a two-dimensional (2D) matrix by recording a window of data and inferring the data to forecast the type of human behavior. Finally, the classification result of the EnsemConvNet method is achieved by combining several classifiers, including product rule method, majority voting, score fusion, and a sum rule called the adoptive weighted method.

In their study, Dai et al.[14] suggested a two-stream attention-based LSTM network based on a visual attention strategy that focuses on the output of each deep feature map and selectively on the most efficient regions of the original input image. A deep feature relation layer is also proposed for altering the DL network parameters based on relation judgments, taking into account the relationships between two deep feature streams. For complex HAR, Khan et al.[15] designed a

26-layer CNN architecture, in which features are derived from the global average pooling and FC layers and combined with the proposed higher entropy-based technique. The researchers also introduced a new FS approach called PDaUM. It is also pertinent to note that some fused CNN characteristics are irrelevant or redundant, resulting in inaccurate predictions in complicated HARs. To manage this issue, Khan et al.[16,17] presented a new HAR method that combines traditional handcrafted features with HOG and deep features, in which human silhouettes are first retrieved in two steps using a saliency-based method. They also proposed an entropy-based FS approach that selects the most discriminative features for classifying M-SVM to deal with the curse of dimension.

Xia et al. LSTM-CNN hybrid was employed for activity recognition. After LSTM gathered the temporal information from sequential multimodal mobile sensor data, CNN retrieved the features. With the UCI-HAR dataset, an $F1$ score of 95.78% (much superior to other based models) was achieved through hyperparameter optimization, such as batch normalization.[15]

Egocentric recognition was used to categorize everyday activities, exercise, ambulation, and office work in the form of a hybrid CNN-LSTM model. In this study, CNN and LSTM were employed in conjunction with egocentric videos and accelerometer to conduct multimodal sensor fusion. However, the results were not as good as they may have been compared with the base model because of a lack of training data.[18]

3 Proposed HAR-HPTDL Model

To detect and classify the occurrence of human activities in an input image, the presented HAR-HPTDL technique implements preprocessing, BiLSTM-based feature extraction, SSA-based hyperparameter optimization, entropy-based feature reduction with Chi square-based feature selection, and SM-based classification modules. Figure 1 shows the overall workflow of the proposed technique, and each module is explained below.

3.1 *Preprocessing*

For the noise reduction process, the WF approach is used. To get as close to the original signals as feasible, WF look for the linear time invariant filter output. The purpose is to minimize the difference between expected output and noise-free signals as much as feasible. WF considers noise to be a generalized stationary process with known second-order statistical features, where the quantity of useable signals is the input.

In the spatial domain, the blurred images are represented by the following equation, in which $f(x, y)$ denotes the input image, $g(x, y)$ indicates the degraded image with few points, spread function $h(x, y)$, and additive noise (x, y):

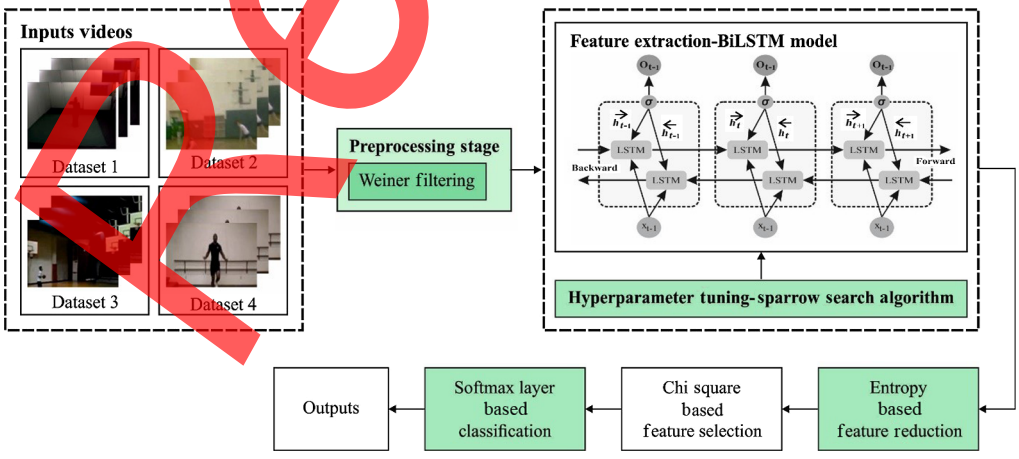

Fig. 1 Overall process of HAR-HPTDL model.

$$g(x,y) = H(x,y)^* f(x,y) + \eta(x,y), \tag{1}$$

where $*$ implies 2D convolutions; $H(x,y)$ represents the blurring function; and additive noise $\eta(x,y)$ refers to uniform noise, Gauss white noise, and so on. The Wiener filter treats noises and images as an arbitrary process and aims to detect an estimation f of the original image $f(x,y)$ where MSE is minimal.[19] The optimization problem is given as

$$\min\ e^2 = E\{(f - \hat{f})^2\}, \tag{2}$$

where E denotes the arithmetical anticipation. In frequency domain, the optimization solutions are expressed as

$$\hat{F}(u,v) = \frac{H^*(u,v)}{(|H(u,v)|^2 + S_\eta(u,v)/S_f(u,v))}, \tag{3}$$

where $H^*(u,v)$ represents the complex conjugate of $H(u,v)$; $S_\eta(u,v)$ indicates the power spectrum of noise; and $S_f(u,v)$ signifies the power spectrum of original image. When $(S_\eta(u,v)/S_f(u,v))$ is higher, the WF gets reduced, and hence, the frequency is neglected.

3.2 *SSA-BiLSTM-Based Feature Extraction*

Sparrow-inspired search algorithms can be an extremely effective optimization tool. Searching is the process of looking into or over something to find or uncover what you are looking for. Sparrow hunt is a term used to describe the practice of gathering food for immediate use or long-term storage. Many species of sparrows live in communities. They can be found all throughout the world, but they like to live around humans. In addition to their value, sparrows have contributed to human culture in a variety of ways. Furthermore, they are and mostly graze on grain seeds or weeds. It is no secret that sparrows are among the most common kinds of year-round residents. Sparrows are highly intelligent and have great memories compared with many other little birds. The simplicity, flexibility, and high efficiency of the SSA are evaluated in engineering applications, and the techniques included in SSA are employed to tackle global optimization challenges. The SSA was inspired by the searching behavior of a sparrow for food. The primary food of the sparrow is grains or weeds. Sparrows are opportunistic, sophisticated feeders who use a range of eating approaches to adapt to the present conditions of their environment and prey. To successfully feed, sparrows use a technique known as foraging, which is described as the acquisition of food through searching, hunting, or gathering. When a bunch of sparrows notices a predator, they all chirp and take flight.

SSA with BiLSTM collects features from the preprocessed image and produces a meaningful set of features using the SSA-BiLSTM model. The RNN has a hard time learning long-term reliance. An RNN based on an LSTM approach is used to tackle a gradient lowering problem. The LSTM approach takes key characteristics from data and stores them for a longer period. As a result, the LSTM method learns the worth of data in terms of removing/keeping it. A, input gate, output gate, and the forgotten gate are common components of the LSTM techniques, which are described in more detail below.

An overview of the structure of long-term memories (LSTM). For example, the input gate $i(t)$ uses previous outputs and current sensor readings to figure out what information is transferred to the memory cell. No need to worry about gate F when it is time to update the memory cell (t). The output gate $o(t)$ selects the data to be transferred to the next time step in the computation.

3.2.1 *Forget gate*

In most cases, the sigmoid function is utilized to determine which of the LSTM memories must be removed. Such decisions basically depend on the values of h_{t-1} and x_t. The result of this gate is f_t, a value between 0 and 1, which indicate the learned value and whole value, respectively. This is how the outcome is judged:

$$f_t = \sigma(W_{f_h}[h_{t-1}], W_{f_x}[x_t], b_f), \tag{4}$$

where b_f indicates a constant known as bias.

3.2.2 *Input gate*

Within the LSTM memory, it now improves the decision of new data. Sigmoid and "tanh" layers are present in this gate,[20] which are assessed as

$$i_t = \sigma(W_{i_h}[h_{t-1}], W_{i_x}[x_t], b_i), \tag{5}$$

$$c_t = \tanh(W_{c_h}[h_{t-1}], W_{c_x}[x_t], b_c), \tag{6}$$

where i_t indicates if the value be either upgraded or not; and c_t signifies a vector of novel candidate value additional to the LSTM memory. A set of these two layers provides an upgrade to the LSTM memory in which the current value is forgotten at the forget gate layer with the multiplications of an older value c_{t-1}, then a new candidate value $i_t * c_t$ is added. The succeeding equation signifies this:

$$c_t = f_t * c_{t-1} + i_t * c_t, \tag{7}$$

whereas f_t represents the outcome of the forget gate which has a value between 0 and 1, respectively, indicating the last value is maintained. Figure 2 shows the framework of Bi-LSTM.

3.2.3 *Output gate*

A portion of the LSTM memory is allotted as a result of the decision being made using the sigmoid layer. The nonlinear tanh function is then created to map values between -1 and 1. After that, the results are multiplied using the sigmoid layer. The equation for calculating the outcome is as follows:

$$o_t = \sigma(W_{o_h}[h_{t-1}], W_{o_x}[x_t], b_o), \tag{8}$$

$$h_t = o_t * \tanh(c_t), \tag{9}$$

where o_t represents the resulting value; and h_t represents a value between -1 and 1.

The foraging and antipredator character of sparrows inspires SSA, which improves the network's overall efficiency by optimizing hyperparameters. Sparrows, which are most reputable for their robust memory capability, exist as producers and scavengers. The former searches for food sources, and the latter collects the food from the producer. SSA can be mathematically formulated considering the foraging behavior of sparrows, whereby virtual sparrows are

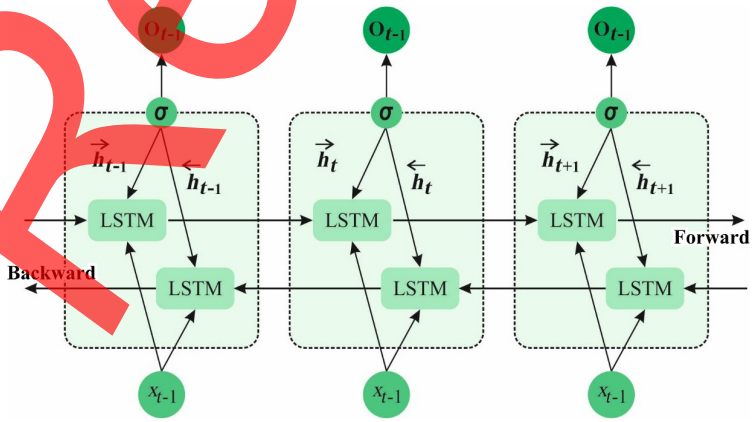

Fig. 2 Structure of Bi-LSTM.

employed to determine the optimum food sources, then the location of the sparrows can be defined as

$$X = \begin{bmatrix} \chi_{1,1} & \chi_{1,2} & \cdots & \cdots & \chi_{1,d} \\ \chi_{2,1} & \chi_{2,2} & \cdots & \cdots & \chi_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & \cdots & x_{n,d} \end{bmatrix}, \tag{10}$$

where n represents the w count and d indicates the way of parameters to be tuned.[21] Therefore, the fitness of the sparrows can be represented as follows:

$$F_X = \begin{bmatrix} f([x_{1,1} & \chi_{1,2} & \cdots & \cdots & \chi_{1,d}]) \\ f([x_{2,1} & \chi_{2,2} & \cdots & \cdots & \chi_{2,d}]) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f([x_{n,1} & x_{n,2} & \cdots & \cdots & x_{n,d}]) \end{bmatrix}. \tag{11}$$

During the searching procedure, the producer with the highest fitness finds the best food source, directs the process of food discovery, and assists the entire population's activities. As a result, the producer has a significant advantage over the scavenger in terms of food identification. Equation (12) can be used to indicate the producer's location:

$$X_{j,j}^{t+1} = \begin{cases} X_{i,j}^{t} \cdot \exp\left(\frac{-i}{\alpha \cdot \text{iter}_{\max}}\right) & \text{if } R_2 < \text{ST} \\ X_{i,j}^{t} + O \cdot L & \text{if } R_2 \geq \text{ST} \end{cases}, \tag{12}$$

where t suggests the current run at $j = 1,2,\ldots,d$; X^t designates the rate of j'th. $\alpha \in (0,1)$ as a random value; $R_2 (R_2 \in [0,1])$ and ST (ST $\in [0.5, 1.0]$) refer to an alarm value and safety threshold value, respectively; Q indicates a random value that employs the simple distribution; and L represents a matrix of $1 \times d$ for every element in j.

In the $R_2 <$ ST scenario, predators are absent, and the producers have a large search area. When $R_2 \geq$ ST, sparrows have found a predator and need to be protected so they can continue to fly in protected locations. In addition, there are not many scavengers following the producers closely. When the producer discovers an ideal food, it leaves the area to compete for it. The meal is yours if you win; else, Eq. (14) is your executed. The updated position of a scavenger can be represented as

$$X_{i,j}^{t+1} = \begin{cases} 0.\exp\left(\frac{X_{\text{worst}}^{t} - X_{i}^{t}}{j^2}\right) & \text{if } i > n/2 \\ X_P^{t+1} + |X_{j,j}^{t} - X_P^{t+1}| \cdot A^{+} \cdot L & \text{otherwise} \end{cases}, \tag{13}$$

where X_P stands for the best possible location for a manufacturer; X_{worst} indicates the current worst-case scenario for the entire world. A is a matrix of $1 \times d$ for a component in 1; and $A^{+} = A^{T}(AA^{T})^{-1}$. When $i > n/2$, the i'th scrounger with unsuccessful fitness is hungry. Consequently, the sparrows that are farther from predator risks have extra lifetime. The essential of the sparrows can be generated arbitrarily in the population. Therefore, it can be numerically defined as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{\text{best}}^{t} + \beta \cdot |X_{i,j}^{t} - X_{\text{best}}^{t}| & \text{if } f_i > f_g \\ X_{i,j}^{t} + K \cdot \left(\frac{|X_{i,j}^{t} - X_{\text{worst}}^{t}|}{(f_j - f_w) + \varepsilon}\right) & \text{if } f_i = f_g \end{cases}, \tag{14}$$

where X_{best} shows the present global optimal place; β implies the step size control parameter that is a normal distribution of arbitrary values with mean value of 0 and variance of 1; $K \in [-1,1]$ is an arbitrary measure; f_i signifies the fitness value of the present sparrow; f_g and f_w are recent

global optimum and least fitness measures, respectively; and ε implies the minimal constant that removes zero-division-error.

In addition, if $f_i > f_g$, then the sparrow is at border of group. X_{est} refers to the place of the center of population that is safe. At this point, $f_j = f_g$ refers to the sparrow in the center of the population, which is aware of a risk and migrates to a nearby edge. K represents the direction in which the sparrow moves and step size control coefficients.

3.3 *Feature Reduction and Feature Selection*

Where XP stands for the best possible location for a manufacturer, X_{worst} indicates the current worst-case scenario for the entire world. As a result, dimension reduction approaches are used to handle this type of problem. Categorization algorithm performance time is reduced after the reduction dimension, but its selective capacity for classifications is kept.[21] Hence, the dimension reduction could enhance predictive accuracy, learning performance, and reduce computation complexity, where XP stands for the best possible location for a manufacturer; X_{worst} indicates the current worst-case scenario for the entire world. The threshold functions are applied for dropping inappropriate features with higher feature values. The computations of entropy technique are determined below.

When using the entropy technique, the resulting fused feature vector $\xi_{Fd}(\text{FV})$ has a dimension of $\xi^{M\times 4096}(\text{FV})$, where M is the total number of tests performed. $\text{FV}(i)$ and $\text{FV}(i, j)$ entropy value together inferred from:

$$E(\text{FV}(i)) = -\sum_i P(\text{Fv}_i)\log_2(P(\text{Fv}_i)), \tag{15}$$

$$E(\text{FV}(i,j)) = -\sum_j P(\text{Fv}_j)\sum_i P(\text{Fv}_i|\text{Fv}_j)\log_2(P(\text{Fv}_i|\text{Fv}_j)), \tag{16}$$

where $P(\text{Fv}_i)$ indicates the previous probability of a fused feature vector; $P(Fv_i|Fv_j)$ implies later probability of each feature ξ_{Fd} (FV); and $E(\text{FV}(i, j))$ denotes an entropy vector. Then, $E(\text{FV}(i, j))$ vectors are arranged in ascending order to estimate likelihood values and elect the MHP features. The MHP values are utilized in features and threshold function, i.e., the feature vector values below MHP are discarded from fused vectors:

$$R(\text{FV}) = \begin{cases} \text{SL} & \text{if } E(\text{FV}(i,j)) \geq P(E(\text{FV})) \\ \text{DC} & \text{otherwise} \end{cases}, \tag{17}$$

where $P(E(\text{FV}))$ represents the MHP value, which is determined as $P(E(\text{FV}))$ = DC and SL represents the elected features, and DC signifies discarded features.[22] Next, χ^2 is executed for selecting an optimal feature. After reduction, a few inappropriate features still exist in the reduced vector (FV). Therefore, a simple χ^2-based FS technique is implemented to measure the degree of associations among features as follows:

$$\chi^2(\text{FV}) = \sum_{i=1}^{K}\sum_{j=1}^{N}\left[\frac{(O(R(\text{FV})_{i,j})) - \mu_{i,j}}{\mu_{i,j}}\right], \tag{18}$$

where $\chi^2(\text{FV})$ represents the selected feature vector that is used in the SM layer.

3.4 *Action Classification*

Finally, the human action classification process is conducted using the SM layer, which is generally the last layer of the DL model. Specifically, the outcome of the convolution and pooling layers are given as input to the SM layer.

Stochastic gradient descent optimization is applied on many iterations and training instances, as well as forwarding propagation, to improve the weights and reduce errors. Every DL algorithm has its own set of parameters and hyperparameters. As an input parameter, the model's

weights. To keep these variables up to date, backpropagation makes use of an optimization process such as gradient descent. In this case, the hyperparameters have been established. In this way, the model's design and learning process are decided by them. Some examples of these parameters are batch size and learning rate as well as the weight decay coefficient.[23] Because DL allows for such model development flexibility, these hyperparameters must be carefully picked for the best outcomes. The SM layer allows the input vector c to be mapped into K classes in an N-dimensional area, as shown in Eq. (19):

$$v_q = \frac{\exp(\theta_q^Z c)}{\sum_{k=1}^{K} \exp(\theta_q^Z c)} (q = 1,2,\ldots,K), \tag{19}$$

where $\theta_k = (\theta_{k1}\theta_{k2}...\theta_{kN}]^Z$ denotes the weights. Here, the SM layer includes the class labels of different actions that exist in the input test images.

4 Performance Validation

The performance of the HAR-HPTDL approach on the HMDB51, UCF101 UCF11, and IXMAS datasets is investigated in this section. The dimensionality reduction procedure of the HAR-HPTDL model performs well in terms of accuracy, as shown in Table 1 and Fig. 3. The results verify that the HAR-HPTDL model can achieve maximum accuracy under distinct values of dimensionality reduction. For instance, with a dimensionality reduction of 128, the HAR-HPTDL model obtained accuracies of 99.68%, 98.15%, 94.87%, 65.98% and on the IXMAS, UCF11, UCF101, and HMDB51 datasets, respectively. Furthermore, with a dimensionality reduction of 1024, the proposed model achieved accuracies of 97.51%, 65.87%, 93.89%, and 99.99% on the UCF11, HMDB51, UCF101, and IXMAS datasets, respectively.

4.1 *Results on UCF11 Dataset*

The UCF11[24] dataset contains a total of 1600 videos gathered from YouTube with 11 actions, whereby each video is related to individual action. Figure 4 displays some video frames from the UCF11 dataset.

Tables 2, 3 and Fig. 5 provide a comparison of the HAR-HPTDL method with established approaches on the UCF11 dataset. According to the experimental results, the dense Traj. model and soft attention approaches performed poorly with accuracies of 0.842 and 0.849, respectively. The BT-LSTM model showed a higher accuracy of 0.853, followed by the DT+BOW+SVM and

Table 1 Entropy-based dimensionality reduction in terms of accuracy on applied dataset.

Dimensionality reduction	UCF11	HMDB51	UCF101	IXMAS
128	98.15	65.98	94.87	99.68
256	99.64	68.67	93.65	99.96
512	97.87	66.09	94.32	99.74
1024	97.51	65.87	93.89	99.99

Table 2 Computation times of existing model and proposed HAR-HPTDL model on applied dataset.

Method	Speed (fps)
STDAN	132
HAR-HPTDL	126

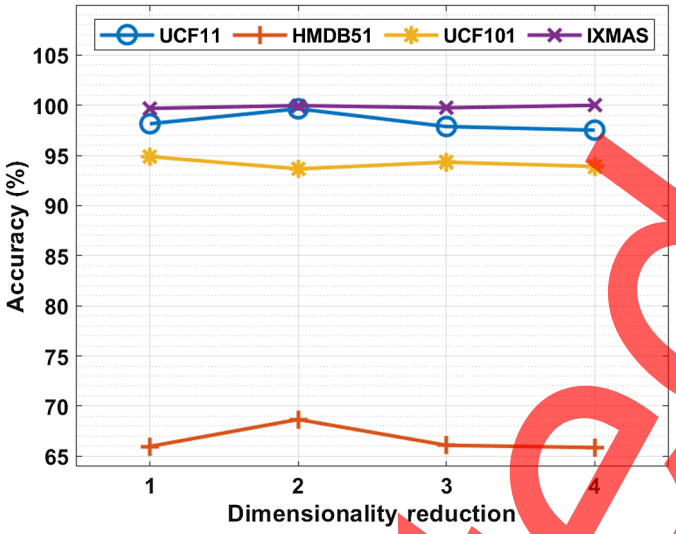

Fig. 3 Results of HAR-HPTDL model in terms of accuracy.

Fig. 4 Samples from UCF11 dataset.

DTAM approaches with moderate accuracies of 0.89 and 0.901, respectively. In addition, the TS-LSTM, TF-VGGNet, and STDAN approaches produced better results, with 0.946, 0.959, and 0.982 accuracy, respectively.[25] However, the new HAR-HPTDL approach outperformed all existing HAR models with an accuracy of 0.996.

4.2 *Results on HMDB51 Dataset*

The HMDB51 action dataset consists of 51 labeled action classes extracted from a variety of sources, including digitized movies and YouTube videos.[26] There are 6766 videos in total, as well as three training or testing splits, in which a total of 3.6K trained video sequences can be

Table 3 Results of existing models and proposed HAR-HPTDL model on UCF11 dataset.

Methods	Accuracy
Dense Traj. model	0.842
DT+BOW+SVM	0.890
TS-LSTM	0.946
BT-LSTM	0.853
DTAM	0.901
Soft attention	0.849
TF-VGGNet	0.959
STDAN	0.982
HAR-HPTDL	0.996

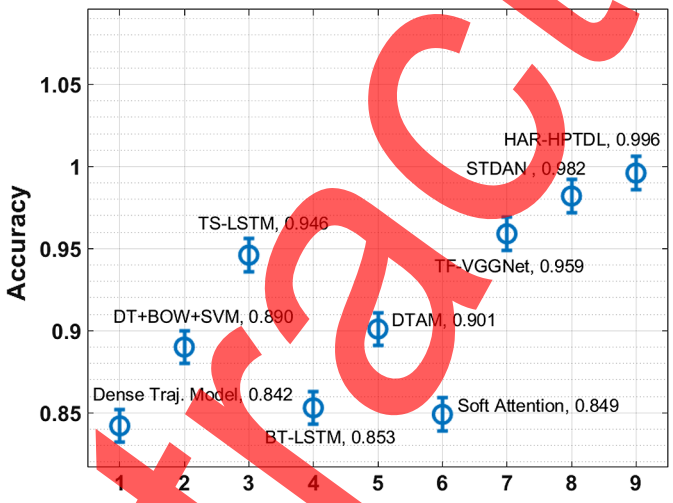

Fig. 5 Accuracy analysis of HAR-HPTDL model on UCF11 dataset.

found across all splits. There are 70 videos to train and 30 videos to test for each class. Figure 6 presents a few representative video frames from the HMDB51 dataset.

Figure 7 shows a detailed comparison of the HAR-HPTDL technique with recent approaches on the HMDB51 dataset.[27] According to the experimental results, the attention-SA, fusion-DIF, and attention-video LSTM approaches performed worst with accuracies of 0.413, 0.428, and 0.433, respectively. The fusion-STR, CNN-I3D, and fusion-MLF techniques achieved better accuracies of 0.454, 0.498, and 0.532, respectively, followed by the attention-RSTAN and attention-residual STAB methods with moderate accuracies with 0.534 and 0.544, respectively. The attention-TCLSTA, TS-AdaScan, attention-MFA, attention-STDAN, TS-ConvNet, attention-JSTA, and TS-STDAN-RGB techniques exhibited even better accuracies of 0.548, 0.549, 0.551, 0.570, 0.594, 0.598, and 0.604, respectively. Overall, the projected HAR-HPTDL technique outperformed all other HAR models with a maximum accuracy of 0.687.

4.3 *Results on UCF101 Dataset*

Its large range of real-world action types makes the UCF101 dataset the most commonly used for action recognition.[28] There are a total of 13,320 videos in the three train-test divisions, each with 9.5K training and 3.7K testing films. An example of one of the UCF101 video frames is shown in

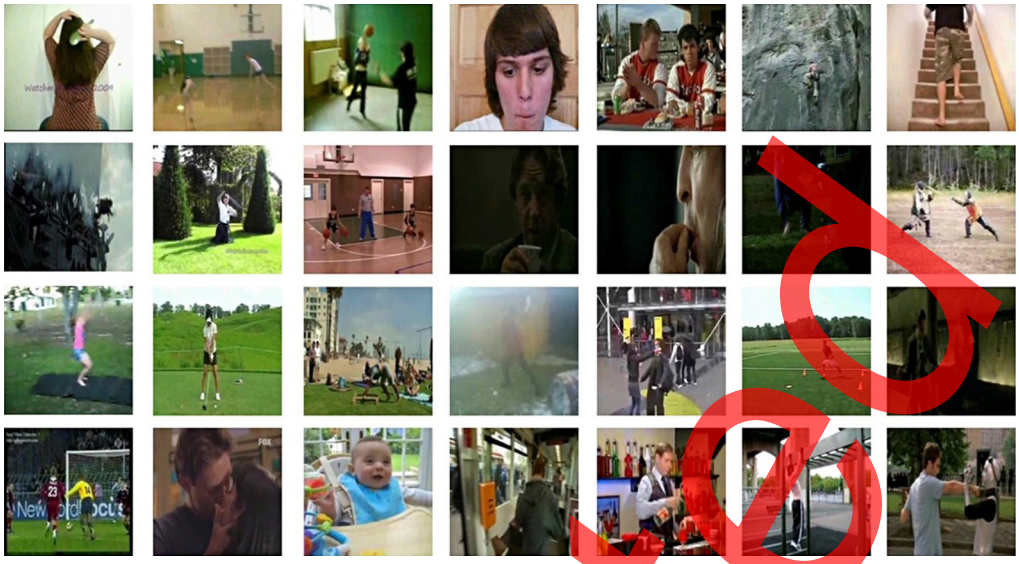

Fig. 6 Samples from HMDB51 dataset.

Table 4 Results of existing models and proposed HAR-HPTDL model on HMDB51 dataset.

Methods	Accuracy
CNN-I3D	0.498
Fusion-MLF	0.532
Fusion-DIF	0.428
Fusion-STR	0.454
Attention-SA	0.413
Attention-video LSTM	0.433
Attention-residual STAB	0.544
Attention-JSTA	0.598
Attention-RSTAN	0.534
Attention-MFA	0.551
Attention-TCLSTA	0.548
Attention-STDAN	0.570
TS-AdaScan	0.549
TS-ConvNet	0.594
TS-STDAN-RGB	0.604
HAR-HPTDL	0.687

Fig. 8. Tables 4, 5, Figs. 8 and 9 show a brief comparison of the HAR-HPTDL method with known methodologies on the UCF101 dataset. Results reveal that the fusion-STR, fusion-DIF, and attention-SA techniques performed poorly with accuracies of 0.758, 0.769, and 0.770, respectively. The attention-video LSTM, attention-RSTAN, and CNN-C3D models showed

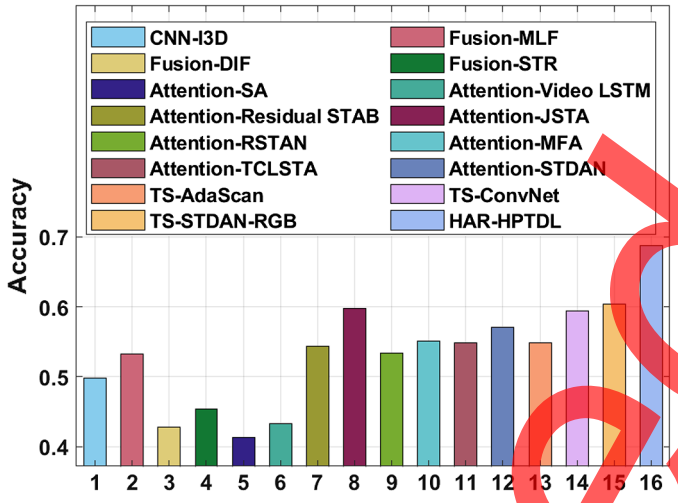

Fig. 7 Accuracy analysis of HAR-HPTDL model on HMDB51 dataset.

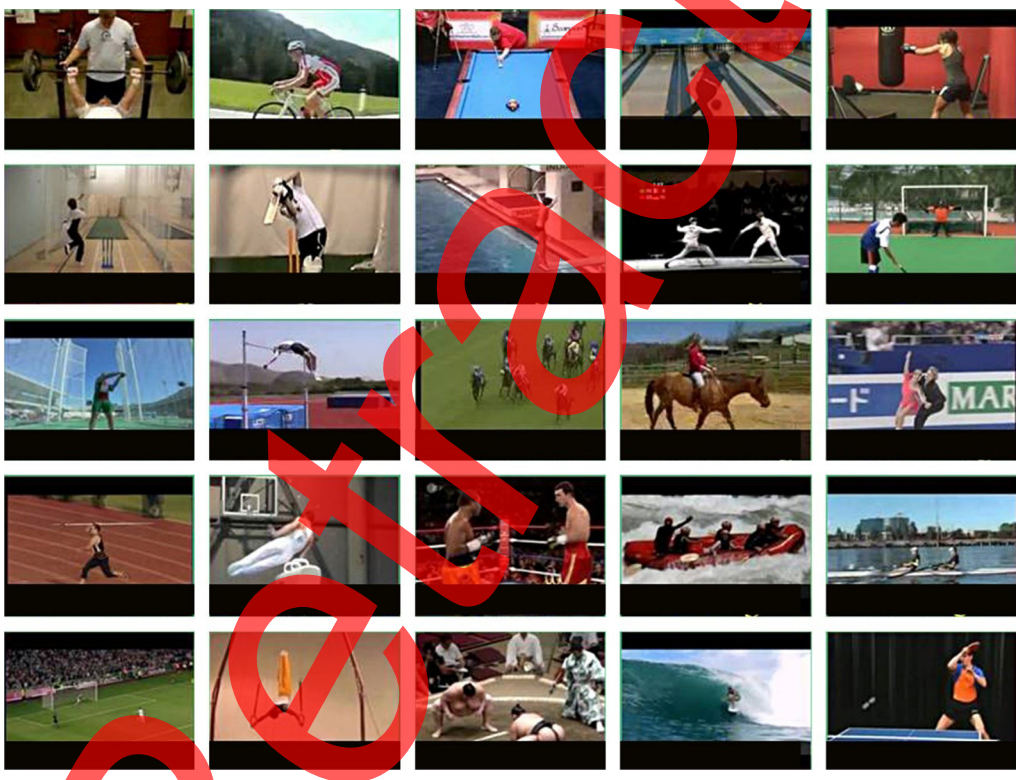

Fig. 8 Samples from UCF101 dataset.

slightly better accuracies of 0.796, 0.802, and 0.823, respectively, while the attention-STAN and CNN-I3D approaches have modest results with 0.828 and 0.845 accuracy, respectively. In addition, the attention-TCLSTA, attention-residual STAB, fusion-MLF, and TS-STDAN-RGB approaches yielded accuracies of 0.859, 0.860, 0865, and 0.873, respectively. Furthermore, the attention-MFA, attention-STDAN, TS-ConvNet, attention-JSTA, TS-AdaScan, and TS-STDAN-RGBD models achieved excellent accuracies of 0.876, 0.877, 0.880, 0.886, 0.894, and 0.910, respectively. Yet, the presented HAR-HPTDL model outperformed the other HAR methodologies with superior accuracy of 0.949.

Table 5 Results of existing HAR models and proposed HAR-HPTDL model on UCF101 dataset.

Methods	Accuracy
CNN-C3D	0.823
CNN-I3D	0.845
Fusion-MLF	0.865
Fusion-DIF	0.769
Fusion-STR	0.758
Attention-SA	0.770
Attention-video LSTM	0.796
Attention-residual STAB	0.860
Attention-JSTA	0.886
Attention-RSTAN	0.802
Attention-STAN	0.828
Attention-MFA	0.876
Attention-TCLSTA	0.859
Attention-STDAN	0.877
TS-AdaScan	0.894
TS-ConvNet	0.880
TS-STDAN-RGB	0.873
TS-STDAN-RGBD	0.910
HAR-HPTDL	0.949

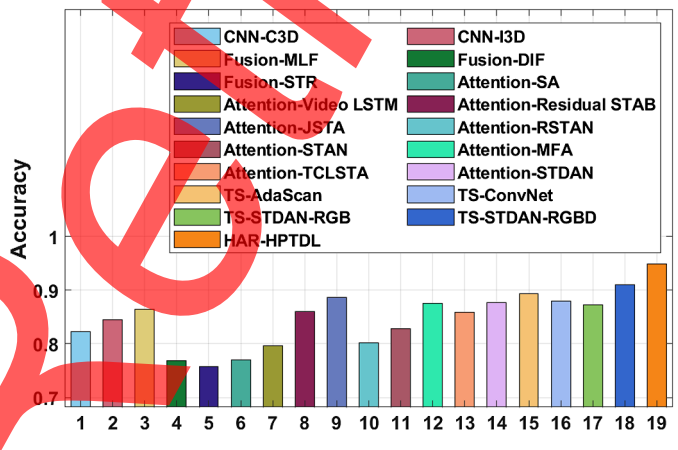

Fig. 9 Accuracy analysis of HAR-HPTDL model on UCF101 dataset.

4.4 *Results on IXMAS Dataset*

It was in 2006 that INRIA built the IXMAS dataset, which comprises 14 action types, such as checking the watch and kicking, waving hands, and punching.[29] The dataset consists of 1148 video 310 sequences at a frame rate of 23 frames per second, with each video being captured by five cameras. Few sample video frames from the IXMAS dataset are depicted in Fig. 10.

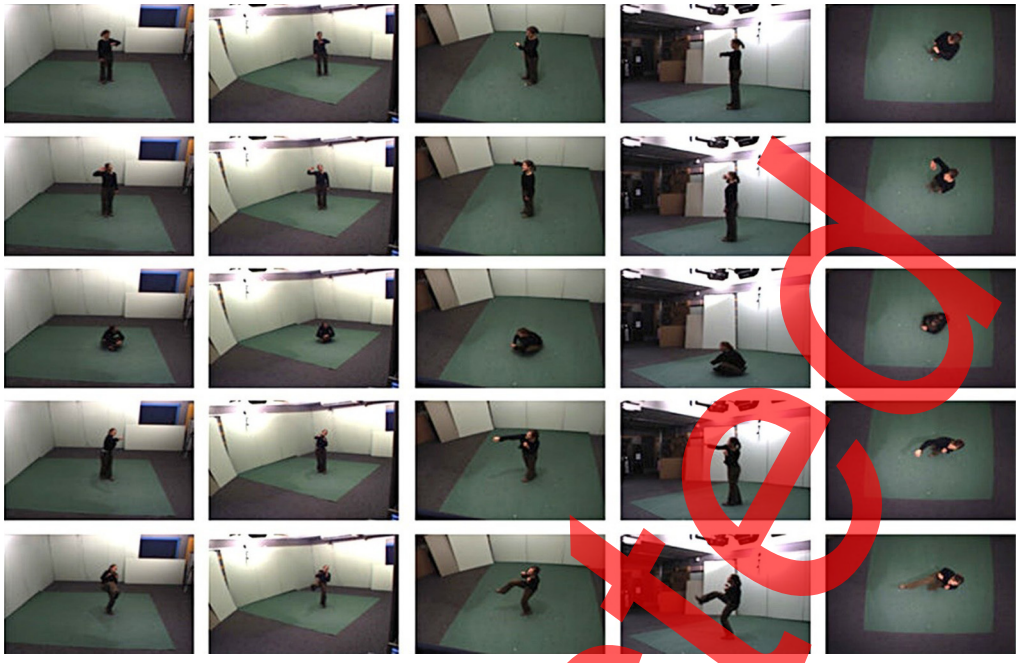

Fig. 10 Samples from IXMAS dataset.

Table 6 Results of existing HAR models and proposed HAR-HPTDL model on IXMAS dataset.

Methods	Accuracy
M-SVM	0.976
L-SVM	0.958
C-SVM	0.990
Q-SVM	0.987
F-KNN	0.997
C-KNN	0.913
B-tree	0.952
W-KNN	0.971
HAR-HPTDL	0.999

A comparative analysis of the HAR-HPTDL technique with other models on the applied IXMAS dataset is demonstrated in Table 6 and Fig. 11. The experimental values suggest that the C-KNN and B-tree methods have the worst performance with a low accuracy of 0.913 and 0.952, respectively. Following that, the L-SVM method yielded somewhat better results (0.958), while the W-KNN and M-SVM approaches yielded more acceptable results (0.971 and 0.976, respectively). Furthermore, the Q-SVM and C-SVM approaches produced better results with 0.987 and 0.990 accuracy, respectively, followed by the F-KNN method with a 0.997 accuracy. The proposed HAR-HPTDL approach, on the other hand, outperformed all previous HAR models with the highest accuracy of 0.999. When examining the comprehensive results analysis, it is clear that the HAR-HPTDL technique is superior to the other current strategies in terms of HAR performance.

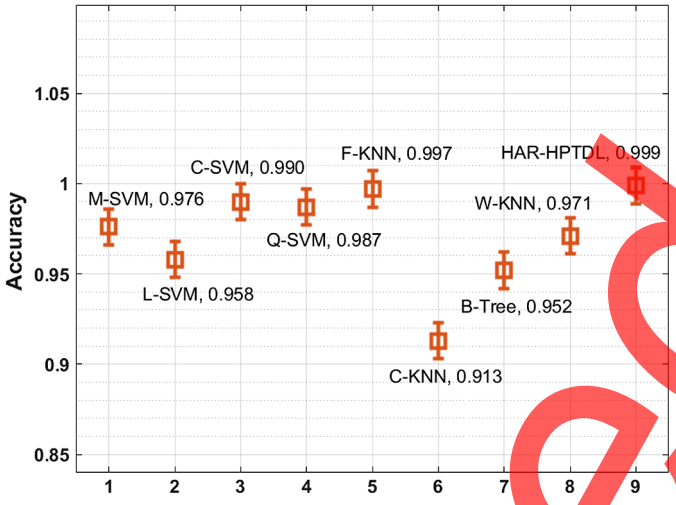

Fig. 11 Accuracy analysis of HAR-HPTDL model on IXMAS dataset.

5 Conclusion

This study proposes a new HAR model using a DL approach, termed the HAR-HPTDL technique, that includes WF-based preprocessing, BiLSTM-based feature extraction, and SSA-based parameter optimization. The design of SSA for hyperparameter tuning of the BiLSTM model enhances the overall HAR performance. Moreover, the proposed models employ entropy-based feature reduction and Chi square feature selection for the detection of discriminant features and removing the repetitive data, respectively. Following a complete simulation investigation on benchmark datasets, the experimental results show that the HAR-HPTDL approach performs well on difficult datasets. However, this is constrained by a lack of subject expertise, and it takes a significant amount of time and resources. This is when deep learning approaches come in handy. Because they do not require human feature engineering on raw data, sensors have shown that deep learning approaches such as convolutional and recurrent neural networks may achieve remarkable results in tough activity recognition tasks. According to our observations and prior findings, an HAR-HPTDL model with the proper handcrafted features and the right hyperparameters can achieve the same level of performance as deep networks on public HAR datasets. In the future, the proposed model can be expanded to include various types of visual modulations, such as RGB and depth data, in the HAR process. The HAR-HPTDL technique can also be used with a variety of sensor data, including wearable accelerometers and gyroscopes.

References

1. H. Eum et al., "Continuous human action recognition using depth-MHI-HOG and a spotter model," *Sensors* **15**, 5197–5227 (2015).
2. L. Chen et al., "Survey of pedestrian action recognition techniques for autonomous driving," *Tsinghua Sci. Technol.* **25**, 458–470 (2020).
3. R. Anand and H. Singh, "Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm," *Measurement* **183**, 109804 (2021).
4. J. A. Ullah et al., "Action recognition in video sequences using deep bidirectional LSTM with CNN features," *IEEE Access* **6**, 1155–1166 (2018).
5. R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identifcation," *Lect. Notes Comput. Sci.* **9912**, 791–808 (2016).
6. F. Serpush and M. Rezaei, "Complex human action recognition using a hierarchical feature reduction and deep learning-based method," *SN Comput. Sci.* **2**(2), 94 (2021).
7. N. Tasnim, M. K. Islam, and J. H. Baek, "Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints," *Appl. Sci.* **11**(6), 2675–2675 (2021).
8. 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.* **32**(4), 447–453 (2020).

9. X. Yu et al., "Deep ensemble learning for human action recognition in still images," *Complexity* **2020**, 1–23 (2020).
10. A. B. Sargano et al., "Human action recognition using transfer learning with deep representations," in *Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 463–469 (2017).
11. E. Mathe et al., "A deep learning approach for human action recognition using skeletal information," in *GeNeDis*, P. Vlamos, ed., pp. 105–114, Springer (2018).
12. Z. Zhang et al., "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," *Neurocomputing* **410**, 304–316 (2020).
13. D. Mukherjee et al., "EnsemConvNet: a deep learning approach for human activity recognition using smartphone sensors for healthcare applications," *Multimedia Tools Appl.* **79**, 31663–31690 (2020).
14. C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Appl. Soft Comput.* **86**, 105820–105820 (2020).
15. S. Satpathy, S. Das, and S. Debbarma, "A new healthcare diagnosis system using an IoT-based fuzzy classifier with FPGA," *J. Supercomput.* **76**(8), 5849–5861 (2020).
16. M. A. Khan et al. "Emergence of a novel coronavirus, severe acute respiratory syndrome coronavirus 2: biology and therapeutic options," *J. Clin. Microbiol.* **58**(5) (2020).
17. M. A. Khan et al., "Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition," *Appl. Soft Comput.* **87**, 105986–105986 (2020).
18. J. Xue and B. Shen, "A novel swarm intelligence optimization approach: sparrow search algorithm," *Syst. Sci. Control Eng.* **8**(1), 22–34 (2020).
19. M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: overview, challenges and the future," in *Classification in BioApps*, N. Dey, A. Ashour, and S. Borra, eds., pp. 323–350, Springer (2018).
20. J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 1996–2003 (2009).
21. H. Kuehne, H. Jhuang, and E. Garrote, "HMDB: a large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2556–2563 (2011).
22. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," CoRR (2012).
23. C. Al-Atroshi et al., "Deep learning-based skin lesion diagnosis model using dermoscopic images," *Intell. Autom. Soft Comput.* **31**(1), 621–634 (2022).
24. P. P. Mathai and C. Karthikeyan, "Deep learning based capsule neural network model for breast cancer diagnosis using mammogram images," *Interdiscip. Sci. Comput. Life Sci.* **14**, 113–129 (2022).
25. D. C. Pretty, R. Cyril, and J. R. Beulah, "An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM," *Concurr. Eng. Res. Appl.* **29**(4), 386–395 (2021).
26. A. Rehman et al., "Automatic visual features for writer identification: a deep learning approach," *IEEE Access* **7**, 17149–17157 (2019).
27. S. I. Kamran et al., "Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease," *Future Gen. Comput. Syst.* **117**, 234–244 (2021).
28. M. Kavitha et al., "Convolutional neural networks-based video reconstruction and computation in digital twins," *Intell. Autom. Soft Comput.* **34**(3), 1571–1586 (2022).
29. K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access* **8**, 56855–56866 (2020).

Deepak Kumar Jain is an assistant professor at the Institute of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China. He received his Bachelor of Engineering degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, India, in 2010, his Master of Technology degree from the Jaypee University of Engineering and Technology, India, in 2012, and his PhD from the Institute of Automation, University of Chinese Academy of Sciences, Beijing, China. His research interests include deep learning, machine learning, pattern recognition, and computer vision.

Xue Liu received her master's degree from North China University of Technology, China. Currently, she works in the School of Basic Course Department, Wuhan Donghu University. Her research interests include higher mathematics education and mathematical modeling.

Subramani Neelakandan (senior member, IEEE) is as an assistant professor in the Department of CSE at R.M.K. Engineering College Chennai. He has 14 years of teaching experience. He received his Bachelor of Engineering in Computer Science and Engineering and his ME degree in computer science and engineering from Anna University Chennai. He has his PhD in information and communication engineering from Anna University. His research interests include data science, machine learning, big data, and cloud computing. He has published more than 30 research papers.

Mohan Prakash (senior member, IEEE) is as an associate professor in the Department of Computer Science and Engineering in VIT University, Vellore, Tamil Nadu, India. He received his Bachelor of Engineering in 2001 from the University of Madras, Master of Engineering (computer science and engineering) in 2007 from Satyabama University, and Doctor of Philosophy (Computer Science and Engineering) in 2014 from Jawaharlal Nehru Technological University Hyderabad. He has more than 20+ years of experience in teaching and research. His area of interest includes data analytics, big data, and machine learning.