Revolutionizing Big Data Assessment for Human Activity Recognition with Metapath Context and Bi-directional Cascade Networks

Submitted: 5th November 2024; accepted: 20th May 2025

Praveen S. Banasode, Sunita Padmannavar

DOI: 10.14313/jamris-2026-047

Abstract:
Human activity recognition focuses on automatically detecting and classifying what individuals are doing, utilizing data gathered from sensors or video feeds. Challenges include interpreting complex activities, handling imbalanced data classes, ensuring privacy in video-based methods, and achieving efficiency with limited computational resources. To solve these issues, this research develops a novel model named the Metapath Contexted Convoluted Bi-directional Cascade Network (MCCBCN), which leverages meta paths to enhance accuracy and robustness in capturing contextual information for human activity recognition. By leveraging both convolutional and bi-directional cascade architectures, MCCBCN makes a significant leap in the field by effectively capturing temporal dependencies. With the addition of the Enhanced Football Team Training Optimization Algorithm (EFTTOA), training efficiency gets a boost, allowing the model to better understand the complex temporal and contextual nuances in human activity data. Additionally, utilizing Large-scale Synthetic Minority Over-Sampling Technique (LSMOST) for dataset expansion ensures a more comprehensive and balanced representation of minority classes, mitigating biases from imbalanced class distributions and improving model generalizability. The proposed method achieved impressive evaluation metrics with an accuracy of 99.50%, F1-score of 99.9%, precision of 99.42%, and recall of 99.3%, outperforming existing techniques. Also, the proposed MCCBCN time complexity is 10.2 seconds. The proposed MCCBCN model effectively addresses key challenges in human activity recognition by enhancing contextual understanding, improving training efficiency, and ensuring robust performance across imbalanced datasets.

Keywords: *Human Activity Recognition, Enhanced Football Team Training Optimization Algorithm, MobileNetV2-Lite, Dig Data, Metapath Contexted Convoluted Bi-directional Cascade Network*

1. Introduction

Human Activity Recognition (HAR) offers valuable insights for numerous sectors, including health monitoring, assisted living, sports and fitness, and surveillance [1]. Video-based HAR has been successfully implemented across many domains. Additionally, recent advancements in sensor technology have made sensors indispensable for analyzing human behavior in healthcare, fitness tracking, behavior investigation, assisted living, and therapy [2]. These systems stand central to pervasive computing, where a multitude of sensors watch the environment and conduct activities [3]. To enable HAR, on-body sensors, namely accelerometers, gyroscopes, and magnetometers, gather motion data that depict the physical movements of persons. These data are important for monitoring medical conditions, fitness, assisted living, and behavioral evaluation [4]. Among the different sensors, wearable sensors show the most interest owing to their omnipresent notion of seamless integration with life. Recognizing human activity is of utmost concern, as it logs behaviors of individuals, generating data that permits computing systems to monitor, analyze, and support daily activities [5]. Systems predominantly fall into two groups: video-based and sensor-based. However, privacy concerns, particularly related to camera installations in personal environments, have prompted the prevalence of sensor-based alternatives [6]. These systems leverage on-body or ambient sensors to precisely detect motion details or record activity patterns while upholding privacy standards. The Big Data Assessment for HAR leverages Deep Learning (DL) systems to examine huge quantities of data for recognizing and categorizing human activities. The advent of big data in HAR has led to the adoption of DL models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [7], facilitating automatic classification from raw sensor data. These approaches facilitate the automatic extraction and classification of attributes from original sensor data. Despite the authority, DL paradigms require big annotated datasets, often not generalizing over various environments due to sensor placements, activity types, and user behaviors. Data augmentation, transfer learning, and adaptive feature learning are considered a few remedies by researchers. Sensor-based HAR systems are a hot topic in health care, wherein they provide a platform for patient monitoring on a continuous and non-intrusive basis. Similarly, these systems measure fitness performance and prevent injuries in fitness applications [8], while, in assisted living, systems respond to caregiver alerts triggered by anomalies in the routines of elderly or disabled persons [9].

To analyze these problems, many hybrid methods in related computer vision problems have been

considered. The combination of robust Principal Component Analysis (PCA) with metaheuristic optimization algorithms aimed at feature extraction and noise reduction, for scenarios involving occlusion or incomplete data, kind of problem faced in HAR, with sensor readings either missing or corrupted [10]. Following this pattern, an architecture combining pretrained detection/meters association/ segmentation modules and lightweight classifiers like K-Nearest Neighbors (KNN) demonstrated an improvement in recognition performance while preserving computational efficiency [11]. Swarm intelligence-based metaheuristics such as Grey Wolf Optimization (GWO) are very useful for feature selection and dimensionality reduction, so systems attend only to the most informative sensor features and end up reducing overfitting [12]. Improvements in skin detection and regional feature localization certainly teach a lot about how to cleanly segment activity-relevant bits from raw input data [13]. Together, these ideas lead to a HAR framework that is more adaptive, interpretable, and resource-efficient to deal with real-world variations [14].

Therefore, to overcome these issues, a novel technique named MCCBCN is introduced, which includes enhanced contextual understanding through rich relational data, improved accuracy by capturing dependencies in both directions, and the ability to handle complex, sequential data more effectively. To enhance the performance of MCCBCN, ETTOA is introduced. These techniques collectively lead to more precise and reliable recognition of human activities, making them highly valuable for analyzing extensive and intricate datasets in HAR.

Contributions

- This research introduces the novel MCCBCN for HAR, effectively capturing contextual information through metapaths to enhance accuracy and robustness. Additionally, this approach of MCCBCN, leveraging both convolutional and bi-directional cascade architectures, contributes significantly to the field.
- The utilization of LSMOST for dataset expansion ensures a more comprehensive and balanced representation of minority classes. By preventing biases caused by class distribution imbalances, it enhances the strength of the model, while empowering it to generalize easily.
- The integration of the MobileNetV2-Lite backbone with Graph Convolutional Networks (MGCN) enables the extraction of human-centric features, while multi-view feature computation captures diverse activity patterns. Selection criteria based on their parameters ensure the inclusion of the most relevant features, enhancing model performance and interpretability.
- The incorporation of compression techniques enhances the FTTOA, named EFTTOA. This enhancement reduces the complexity and cost associated with integrating advanced pose estimation techniques. This algorithm has the ability to provide precise activity recognition while minimizing computational resources, making it efficient for applications.

The structure of the work is delineated as follows: Section 2 offers a literature survey, presenting a summary of earlier research and findings in the field. Section 3 elucidates the proposed method, delineating the approaches utilized in the research. Section 4 illustrates the outcomes and discussion, analyzing the outcomes of the methodology and deliberating on their implications. Section 5 provides the conclusion.

2. Literature Survey

This section delves into recent literature on HAR, expanding on prior research efforts. Numerous studies have emerged, and this discussion focuses on showcasing current developments in the field.

Nguyen et al. [10] developed the Triple-feature Double-motion Network (TD-Net), an enhanced iteration of the Double-feature Double-motion Network (DD-Net) featuring an additional branch. This new branch incorporated standardized joint coordinates to enhance spatial information. Zhang et al. [11] incorporated the Discrete Wavelet Transform (DWT) method into the discrete transform system to enhance the expressive features of human actions. By investigating pre-trained dual-stream Convolutional Neural Networks (CNN)- Recurrent Neural Networks (CNN-RNN), learned features combined with handcrafted ones to address the severe demands within the domain. Challa et al. [12] developed a vigorous classification approach for HAR employing data from wearable sensors, employing a hybrid technique that combines CNN and Bidirectional Long Short-Term Memory (BiLSTM). This multibranch CNN-BiLSTM network is intended to extract features directly from raw sensor data with negligible preprocessing required. Luptakova et al. [13] developed the HAR transformer model, originally designed for repurposing for analyzing time-series motion signals. Leveraging its self-attention mechanism, the transformer captured individual dependencies among signal values within the time series. Khatun et al. [14] developed a hybrid DL that coupled CNN and LSTM (CNN-LSTM) and enhanced it with a self-attention algorithm to boost the model's predictive abilities. In addition to the database, the approach was assessed using benchmark databases like MHEALTH and UCI-HAR to validate its comparative performance. Shanableh [15] developed a new method for extracting features for HAR in videos, utilizing video encoding principles like motion compensation and attribute variables based on coding. These features were then used with DL for model classification. Ahmad [16] suggested HAR utilizing deep CNN features.

Subsequently, the critical features within deep representations were selected, and these selected

features were fed to learn temporal dynamics. Hassan et al. [17] presented a dynamic HAR technique based on a Deep BiLSTM model and feature extraction via pre-trained transfer learning. The approach initially employs CNN models, specifically MobileNetV2, to extract deep-level features from video frames. Liu et al. [18] developed a new device-free method called TransTM (Time-streaming Multiscale Transformer). This model harnesses the data-fitting capabilities of transformers to directly use raw RFID RSSI data as input, bypassing the need for pre-processing. Cob-Parro et al. [19] presented a method that was devised for detecting people and recognizing their actions in natural settings, tailored and optimized for real-time operation on edge devices. HAR was implemented using an RNN, specifically an LSTM. The LSTM received as input a lightweight, ad hoc feature vector extracted from the bounding box around every detected individual in video surveillance footage. Table 1 presents a comparative analysis of HAR against established methods.

2.1. Problem statement

In recent years, researchers have increasingly employed DL-based frameworks for HAR. However, persistent issues within these frameworks highlight various limitations, necessitating the development of improved approaches. Previous methods have identified challenges like deficiencies in activity recognition, low accuracy, performance disparities across various datasets, higher time, and lower convergence. These shortcomings underscore the motivation for the MCCBCN methodology in HAR. The primary goal is to address these issues and enhance overall HAR performance by leveraging optimization techniques, aiming for a more robust and effective solution that produces precise and contextually relevant recognition results.

3. Proposed Methodology for Human Activity Recognition

In this research, four human action datasets are used: JHMDB, UCF11, UCF YouTube Action, and HMDB51.These datasets underwent expansion using the LSMOST to generate synthetic samples for minority classes, creating extensive and balanced data to mitigate biases during model training. Features were extracted using a MobileNet v2-Lite backbone with GCN, followed by computing multi-view features from gradients in both horizontal and vertical directions, with features oriented vertically.

Fig. 1 illustrates the block diagram of the introduced work. These extracted features were then combined and selected based on relative entropy, mutual information, and the SCC via a maximum probability-based threshold function. In the next phase, the selected features were fed into the MCCBCN to capture time-based, spatial, and behavioral relationships in video data streams. The model's accuracy and robustness were further enhanced with the integration of the EFTTOA.

3.1. Dataset collection

Input images are sourced from four diverse datasets like JHMDB [20], UCF11 [21], UCF YouTube action [22], and HMDB51 [23], comprising 928 video sequences categorized into 21 action classes. It includes physically marked skeleton data extracted from RGB videos and is partitioned into random train/test splits. The UCF11 dataset, also known as the YouTube action dataset, includes 1,160 videos in 11 action classes and presents challenges due to diverse conditions in the clips. The UCF YouTube action dataset, with 11 action classes like horse riding and biking, is challenging due to variations in camera movement, viewpoint changes, diverse poses, intricate backgrounds, and was compiled by 25 different groups. The HMDB51 dataset contains 6,849 videos in 51 classes, each with at least 101 videos, sourced from movie clips and YouTube.

3.2. Big data assessment using local linear synthetic minority over-sampling technique

After data collection, the LSMOST is applied to address class imbalance in the dataset. This involves identifying minority class instances that are underrepresented compared to the majority classes. LSMOST then generates synthetic samples specifically for these minority classes by including existing instances and their nearest neighbors in a locally linear manner. This process aims to raise the number of minority class samples, effectively balancing the class distribution in the dataset. By doing so, LSMOST ensures that all classes are adequately represented throughout the training phase of machine learning models, thereby mitigating the potential bias towards majority classes and improving the model's overall performance and accuracy. This establishes a clear sequence in the process where the results from the big data using LSMOST are directed towards the feature extraction [24].

3.3. MobileNetV2-lite backbone with graph convolutional networks for feature extraction

The collected dig data undergoes feature extraction using a MobileNetV2-lite backbone with Graph Convolutional Networks (MGCN). The MobileNetv2-Lite is a lightweight and efficient feature extractor for images or videos. The MobileNetv2-Lite produces small but rich features that capture important properties. These features are then fused with the MGCNs, which are capable of modeling relationships in graph-structured data. The proposed method rests on leveraging the efficiency of MobileNetv2-Lite as a feature extractor and the relational learning capacity of MGCNs to create a rich and expressive overall feature representation or embedding. Figure 2 displays the MobileNetV2 with the MGCN model.

During this stage, MGCNs further process the features generated from MobileNetv2-Lite, supplementing the feature representation with contextual data important for downstream processing. Finally,

Table 1. Comparison of the existing work with HAR

Ref	Methods	Pros	Cons	Performance measures used
[10]	TD-Net	The method was effective in both dimensions, delivering strong performance on extensive datasets.	It failed to address the computational complexity associated with implementation, limiting its practical application in scenarios requiring immediate processing.	Accuracy-79.3% F1-score-64.15% Training time-0.5ms Inference time-0.05ms Params-1.88M GFLOPs-0.028
[11]	CNN-RNN	The DT model enhances the descriptive human action features by integrating the DWT technique. This approach improves the performance of action recognition by leveraging both engineered features and learned ones, thereby harnessing the advantages of both methodologies.	Absence of temporal domain diversity in terms of time scales.	Accuracy-95.68
[12]	CNN-BiLSTM	The approach was intended to seize both short-range and extended dependencies within sequential data, ensuring a comprehensive understanding of temporal relationships across various time scales.	Inability to incorporate contextual information from multiple branches into the model architecture hindered its capability to effectively seize nuanced outlines in the data.	Accuracy-96.37 F1-score-96.31 Training time-53.68 Params-631,014
[13]	HAR transformer	Greater temporal spacing between features in the time series provided enhanced contextual comprehension across extended sequences.	Lack of testing the modified transformer on a prolonged database, preferably integrating diverse sensor data, limited a comprehensive evaluation.	Accuracy-99.2 Precision-99.2 Recall-99
[14]	CNN-LSTM	The recognition accuracy was improved by leveraging user similarity and weighting training data, allowing for a more refined learning process that accounted for individual user characteristics and preferences.	It failed to focus on the advances observed in wearable and phone-based tracking systems.	Accuracy-98.76 Params-634,052 Recall-99.93
[15]	ViCo-MoCo-DL	This method effectively utilizes video encoding and motion estimation techniques, combined with DL, for enhanced HAR.	Absence of evaluation across diverse datasets limits the generalizability of the findings.	Accuracy-97%
[16]	CNN-Bi-GRU	This technique leverages DL through CNN and Bi-GRUs for improved HAR performance.	Lack of comprehensive feature selection techniques, theoretically leading to suboptimal performance and increased computational complexity.	Accuracy-93.38 Average time-7.15s
[17]	Deep BiLSTM	The utilization of the model improved through transfer-learning-based attribute extraction for dynamic HAR.	Inability to evaluate across diverse datasets, potentially limiting the generalizability of the model's performance.	Accuracy-99.2%
[18]	TransTM	By utilizing a time-streaming multiscale transformer for device-free HAR.	Lack of extensive validation across diverse scenarios, potentially limiting the generalizability of its performance outside controlled environments.	Accuracy-99.1% F1-score-99.1% FLOPs-27.98M Params-1.82M
[19]	RNN-LSTM	Its capability to perform DL based HAR directly on edge devices.	Inability to constrain resources in edge environments, potentially limiting the model's performance or scalability on such platforms.	Accuracy-99.31% Recall-99.09% F1-score-99.06% Time-0.298ms

the combined approach produces a final feature representation that incorporates the original features obtained from MobileNetv2-Lite into the relational information stored in MGCNs. The feature extraction method with a MobileNetv2-Lite backbone and MGCNs enables efficient processing of intricate graph-structured data, providing important insights for comprehension and decision-making in different

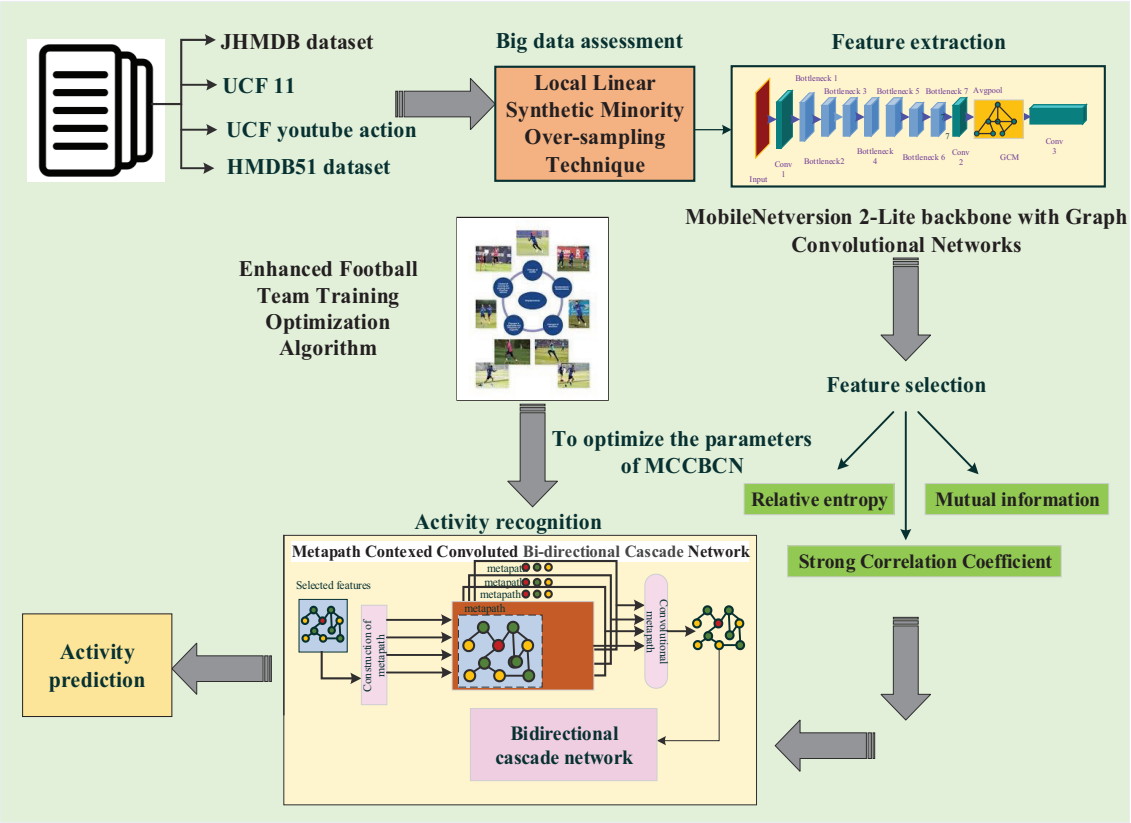

Figure 1. Workflow of the proposed work

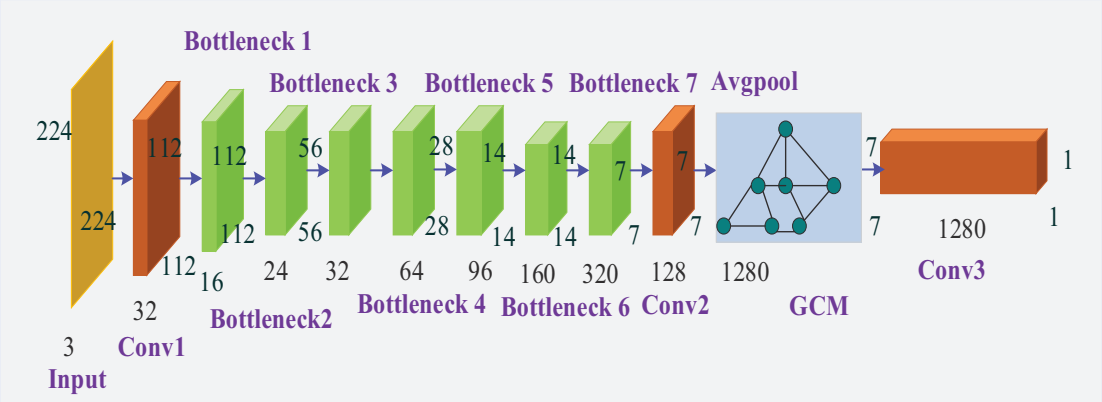

Figure 2. MobileNetV2 with GCN model

application fields. The extracted features are then transferred to feature selection to pull out the more important features [25, 26].

3.4. Feature selection using parameters

The procedure for identifying the most pertinent features involves combining all extracted features. This selection process relies on three key parameters: relative entropy, mutual information, and the SCC, which are assessed using a maximum probability-based threshold function.

- Relative entropy

Relative entropy measures the difference between two probability distributions. In this feature selection, it evaluates the information gain provided by each feature compared to a baseline distribution. Features with higher relative entropy values are considered more relevant, as they contribute significantly to distinguishing between different categories in the dataset. The relative entropy is expressed in Eqn. (1):

$$G_{LK}\left(Q \| P\right) = \Sigma_x P\left(x\right) \log\left(\frac{P\left(x\right)}{Q\left(x\right)}\right) \qquad (1)$$

where, $G_{LK}\ (Q\|P)$ is the relative entropy between distributions of Q and P, $P(x)$ is the probability of a feature x occurring in the database and $Q(x)$ is the probability of a feature x occurring in the baseline distribution. Higher relative entropy values indicate features that contribute more to discriminating between different categories in the dataset.

- Mutual information

It quantifies the dependency among each feature a and the target variable y (e.g., action labels in human motion data). It measures how much information one variable delivers regarding another variable. Therefore, the expression for mutual information is given in Eqn. (2):

$$I(A;B) = \sum_{A}\sum_{B} P(A,B)\ \log\left(\frac{P(A,B)}{P(A)\,P(B)}\right) \qquad (2)$$

where, $I(A;B)$ is the mutual information between features A and the target variable B, Higher mutual information values indicate stronger associations among features and the target variable.

- Strong correlation coefficient

The SCC evaluates the linear correlation strength between pairs of features. It identifies features that exhibit strong correlations with each other, which indicate redundancy in the feature set. To avoid redundancy and overfitting, a maximum probability-based threshold function is applied to the SCC values. The Pearson correlation coefficient formula for SCC is expressed in Eqn. (3):

$$\rho\,(A,B) = \frac{Cov\,(A,B)}{\sigma_a \sigma_b} \qquad (3)$$

where, $\rho\,(A,B)$ is the Pearson correlation coefficient between features A and B, is the covariance between features A and B, σ_a and σ_b are the standard deviations of features A and B, respectively. By setting a threshold τ, features with SCC values above this threshold are selected, as expressed in Eqn. (4):

$$Selected\ features = \{X_i |\, \rho\,(A,B) > \tau\} \qquad (4)$$

By considering these parameters and their corresponding equations, the feature selection process ensures that the selected attributes are extremely informative, non-redundant, and relevant for the given DL task. The chosen attributes are then moved to MCCBCN for activity recognition.

3.5. Activity recognition using metapath contexted convoluted bi-directional cascade network

During the HAR process, the extracted features, characterized by fixed sequence lengths, are forwarded to the proposed method, MCCBCN. This phase is critical for capturing the complex dependencies that characterize human activities in video data streams. The MCCBCN is designed to effectively model and analyze these dependencies by leveraging a combination of advanced techniques. In metapath contextual learning, metapaths are sequences of actions or poses over time, capturing the semantic relationships between different activities.

The MCCBCN model is proposed to boost human activity recognition by combining graph-based context modeling and deep sequence learning. It utilizes metapath-guided context extraction to capture semantic high-level interactions between heterogeneous activity features, allowing the model to identify minute interdependencies between actions. The convolutional layers are nonlocal spatial feature extractors from input streams (e.g., sensor or video frames), and the bidirectional cascade modules yield efficient temporal aggregation by progressively combining forward and backward temporal flows. The cascade mechanism facilitates hierarchical fusion, in which outputs from previous layers are iteratively refined by later-stage features to ensure deep temporal coherence. The model's architecture prioritizes cross-context attention propagation and multi-level abstraction to be able to well capture both long-range activity dependencies and short-term transitions in an end-to-end trainable, scalable manner.

Mathematically, a metapath ρ in a diverse evidence denoted as a sequence of node types $v_1, v_2, ..., v_k$ and edge types $e_1, e_2, ..., e_{k-1}$. The HAR is expressed in Eqn. (5):

$$\rho = \left(v_{sitting}, e_{transition}, v_{s\,\tan ding}, e_{transition}, v_{walking}\right) \qquad (5)$$

By utilizing these metapaths, the network understands and controls the temporal context in which each activity occurs. The contextual information is crucial for distinguishing activities that appear similar in isolated frames but differ in their temporal sequence. It also involves convolutional processing, where GCNs are applied to nodes and edges defined by the metapaths. In HAR, nodes represent individual frames or body joints, and edges represent transitions between these nodes over time. The convolution operation in a GCN for a node i is expressed in Eqn. (6):

$$h_i^{(l+1)} = \sigma\left(\Sigma_{j\in N(i)} \frac{1}{\sqrt{d_i d_j}} W^{(l)} h_j^{(l)}\right) \qquad (6)$$

where, $h_j^{(l)}$ is the attribute vector of the node i at layer l, $N(i)$ is the set of neighbors of the node i, d_i and d_j are the degrees of the nodes of i and j, $W^{(l)}$ is the weight matrix at layer l, σ is an activation function. This operation aggregates features from neighboring nodes, capturing spatial dependencies and learning robust features representing spatial configurations and movement patterns. Bi-directional learning processes data in both forward and backward directions along the metapaths, analyzing sequences of actions from the past to the future and vice versa.

In the bi-directional learning method, the hidden state h_t at time t is a mixture of the forward hidden state $\vec{h}_t$ and backward hidden state $\overleftarrow{h}_t$ which is expressed in Eqn. (7):

$$h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right] \tag{7}$$

The forward and backward hidden states are expressed as in Eqns. (8) and (9):

$$\vec{h}_t = f\left(W_x x_t + W_h \vec{h}_{t-1} + b\right) \tag{8}$$

$$\overleftarrow{h}_t = f\left(W_x x_t + W_h \overleftarrow{h}_{t+1} + b\right) \tag{9}$$

where, x_t is the input at time t, W_x and W_h are weight matrices, b is the bias vector, and f is the activation function. Bi-directional learning captures dependencies that span across different time frames, providing a more comprehensive understanding of the activity. MCCBCN employs a multi-layered approach, where the outcome from one layer is used as the input for the subsequent layer. This cascading structure allows for progressive refinement of the learned features. The input of the l^{th} layer $H^{(l)}$ is expressed as in Eqn. (10):

$$H^{(l+1)} = \sigma\left(W^{(l)} H^{(l)} + b^{(l)}\right) \tag{10}$$

where, $H^{(l+1)}$ is the output of the layer l, $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector. In HAR, the first layer detects basic actions like "aising hand" or "moving leg," the next layer recognizes more complex sequences like "waving" or "walking," and the final layer identifies overall activities like "greeting" or "exercising." This multi-stage processing builds increasingly complex and discriminative representations. MCCBCN is a powerful method for HAR that leverages metapath-based contextual learning, convolutional processing, bidirectional learning, and cascading networks. By capturing temporal, spatial, and behavioral dependencies, MCCBCN accurately recognizes and differentiates various activities, even in complex environments. The network's capability to learn from metapaths, aggregate spatial features, process bidirectional sequences, and refine features through cascading layers results in robust HAR [27, 28].

3.6. Enhanced football team training optimization algorithm to boost the performance of MCCBCN

The introduction of EFTTOA aims to enhance the performance of MCCBCN in HAR. The EFTTOA, implemented as the training policy for MCCBCN, further increases the adaptability and convergence efficiency of the model. EFTTOA mimics the dynamics of a football team, with players (solution agents) assigned roles and strategies through cooperative learning and competitive performance. Agents experience repeated formation change, role exchange, and ability adaptation, allowing for effective global exploration in the parameter space. Adaptive balance between exploration (searching new solutions) and exploitation (optimizing promising regions) in EFTTOA prevents premature convergence and encourages optimal hyperparameter tuning. It works especially well with the high-dimensional optimization problems posed by the MCCBCN's deep structure, providing faster convergence rates and better generalization without degrading model stability or increasing computational costs. EFTTOA is a modification of FTTOA [29] that leverages principles from football team training methodologies to optimize the training process of MCCBCN and improve its effectiveness in identifying human emotions from video data streams. To address the challenges of complexity and cost associated with integrating advanced pose estimation techniques in FTTOA, implementing model compression techniques is a suitable enhancement. The fitness function is expressed as in Eqn. (11):

$$Fitnessfunction = \min\{W^{(l)}\} \tag{11}$$

By compressing DL models like Convolutional Pose Machines (CPMs), the computational complexity and resource requirements are significantly reduced without compromising performance, which makes the models more efficient and cost-effective to implement. The mathematical model compression is expressed as in Eqn. (12):

$$\text{Compressed model} = \arg\min_{M'} L(M', D) + \lambda \cdot \text{Compression\_loss}(M') \tag{12}$$

where L represents the original loss function, D is the training dataset, M' is the compressed model, and $Compression\_loss$ is a regularization term that penalizes complexity. The hyperparameter λ controls the trade-off between accuracy and compression. This enhancement reduces the computational burden and resource requirements, making the integration of pose estimation models more feasible for football teams and organizations. Table 2 displays the pseudo code of EFTTO.

Overall, the activity recognition utilizes an MCCBCN, which is designed to effectively capture and analyze complex relational data to understand activity contexts more accurately. This network structure processes data in both directions across several layers. This algorithm optimizes the training process by fine-tuning the network's parameters, resulting in more accurate and efficient activity recognition. This integrated method is designed for applications that need detailed and accurate activity recognition.

4. Results and Discussions

In this section, the efficiency of the introduced framework for HAR is estimated by relating it to existing methods across four challenging datasets, such as JHMDB [20], UCF 11 [21], UCF YouTube Action [22], and HMDB51 [23]. This research is implemented using Python. The parameter setting of the introduced approach is explained as follows: In this training setup, the maximum detection size for the input data is set

Table 2. Pseudo code of EFTTO

Initialize FTTA
CLC
// Function parameters objectives set
Set Max_{gen} *= 500 // Iterations numbers*
Set n_{pop} *= 50 // Population size*
Set P_{study} *= 0.2 // Probability of study*
Set P_{comm} *= 0.2 // Probability of communication*
Set P_{error} *= 0.001 // Probability of error*
// Initialization
For I = 1 to Max_{gen} *Do*
// Initialization loop starts here
For each i from 1 to n_{pop} *Do*
The fitness value of each football player is calculated, which is expressed in Eqn. (11)*:*
End for
Find the finest player & the worst player
// Collective Training Phase begins
Collective training:

$$H_{i,j}^{k}new = \begin{cases} H_{i,j}^{k}old + rand * \left(H_{best,j}^{k} - H_{i,j}^{k}old\right) \\ H_{i,j}^{k}old + rand_1 * \left(H_{best,j}^{k} - H_{i,j}^{k}old\right) - rand_2 * \left(H_{worst,j}^{k} - H_{i,j}^{k}old\right) \\ H_{i,j}^{k}old + rand * \left(H_{best,j}^{k} - H_{worst,j}^{k}\right) \\ H_{i,j}^{k}old * (1 + K\,(t)) \end{cases} \tag{13}$$

$$H_{i,j}^{k,team_1}new = \begin{cases} H_{best,j}^{k,team_1} if\ rand \le p_{study} \\ H_{random,j}^{k,team_1} if\ rand \le p_{study} \\ H_{random,j*(1+randn)\ if\ rand \le P_{comm}}^{k,team_1} \end{cases} \tag{14}$$

$$H_{i,j}^{k,team_1}new = H_{random_1,random_2}^{k,team_1} if\ rand \le p_{error} \tag{15}$$

For every $i : 1 \le i \le n_{\mathrm{Pop}}$ *do*
End for
Update the players
Find the best player
Additional individualized training:

$$H_{best}^{k}new = \begin{cases} H_{best}^{k}old * \left(1 + \left(\frac{1}{k}\right) * gauss + \frac{1}{k} * cauchy\right) if\ H\left(H_{best}^{k}new\right) < H\left(H_{best}^{k}old\right) \\ H_{best}^{k}old\ if\ H\left(H_{best}^{k}new\right) > H\left(H_{best}^{k}old\right) \end{cases} \tag{16}$$

End for
$T = T + 1$
Exit the while loop
Exit the for loop

to 256 × 256 pixels. The model is trained over 100 epochs using the EFTTOA optimizer. Each epoch consists of 100 iterations with a batch size of 64, meaning 64 models are handled before the approach's parameters are reorganized. The learning rate is set to 0.0001. Additionally, a dropout rate of 0.03 is applied to avoid overfitting by randomly dropping 3% of the neurons during each iteration.

Figures 3 (a), (b), (c), and (d) display the confusion matrix of the corresponding dataset. This confusion matrix is utilized to assess the performance of a recognition model by exhibiting the counts of actual versus predicted values.

Fig. 4 shows the class-wise accuracies of the corresponding datasets.

4.1. Qualitative analysis

Figures 5 (a), (b), (c), and (d) have four videos selected from the test set to evaluate and analyze the experimental quality of HAR generated by the proposed method, comparing it with the corresponding datasets. In the first, a girl is seen combing her hair. Upon examining all tags for HAR, it was noted that this specific activity is absent from the training data.

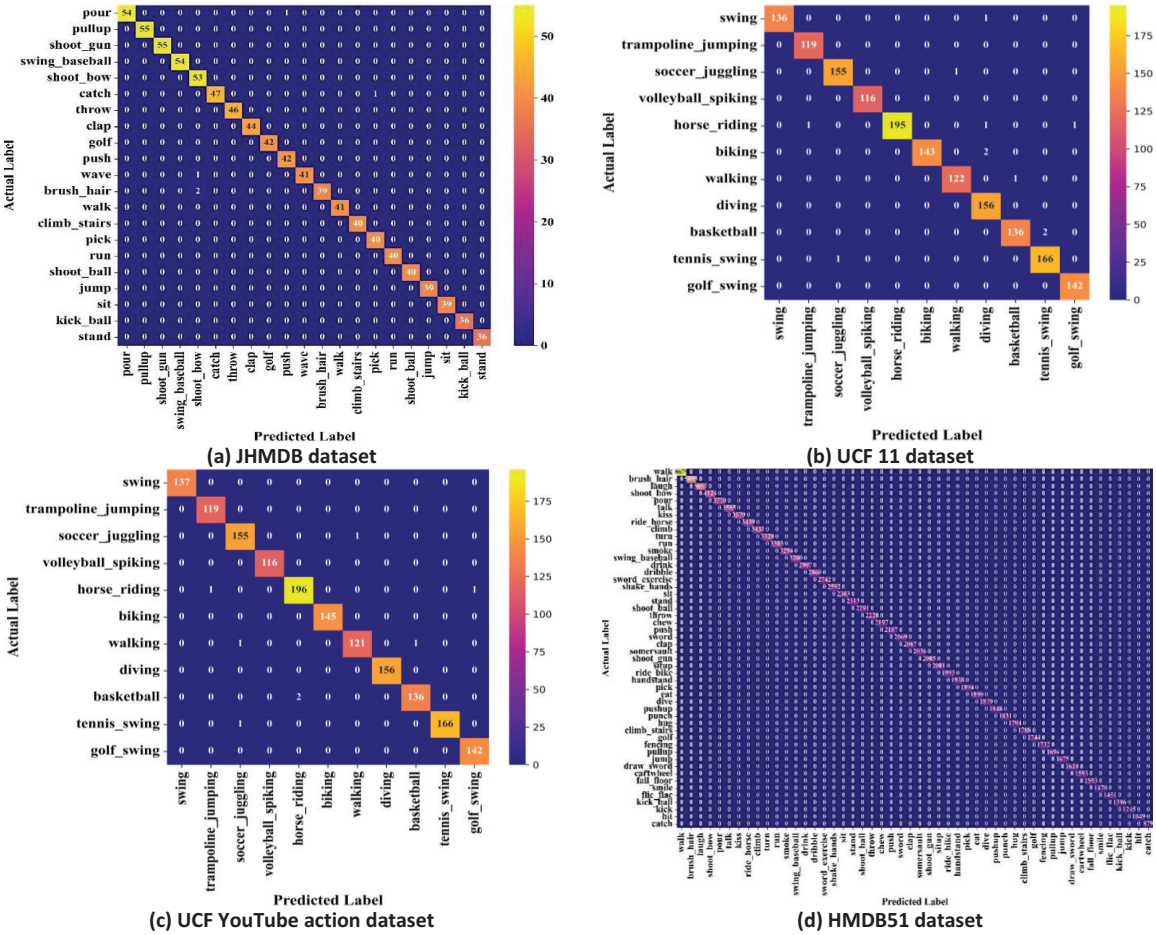

Figure 3. Confusion matrix of the corresponding datasets

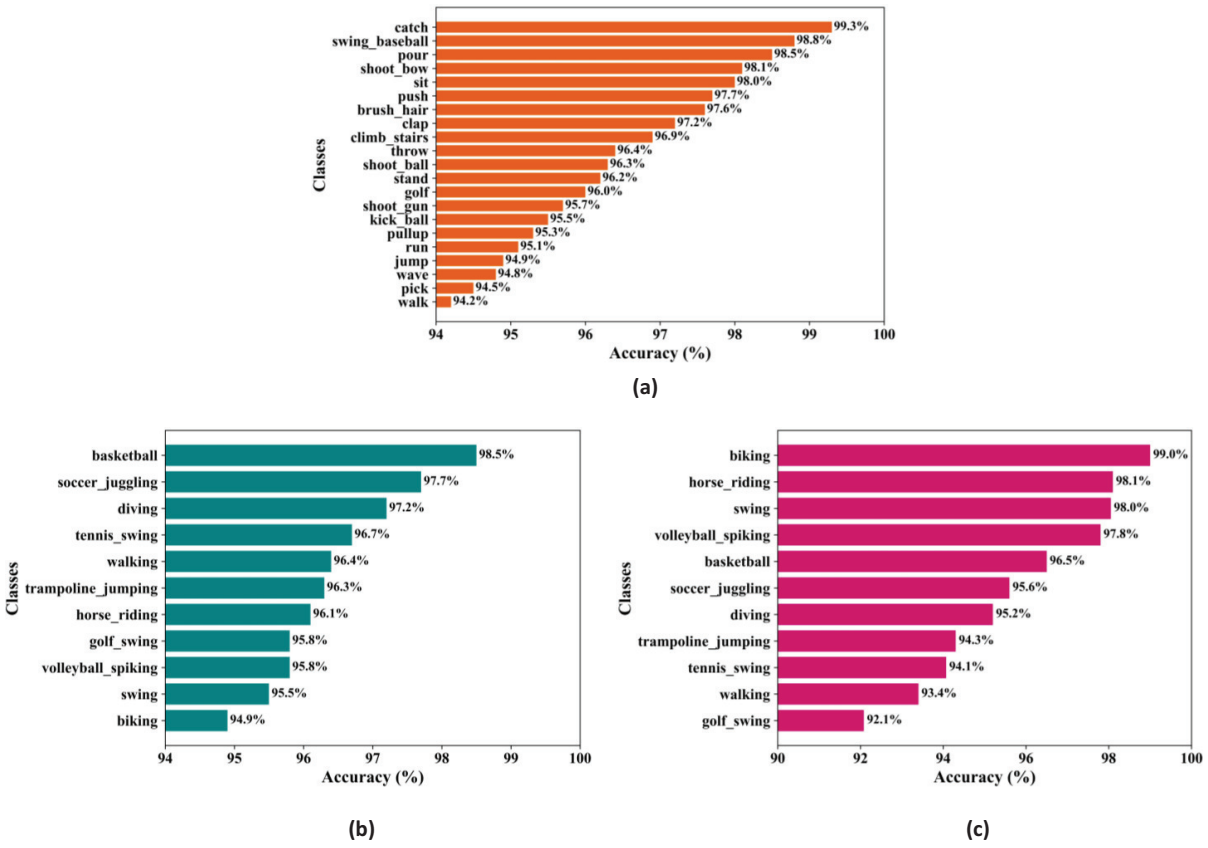

Figure 4. (a, b and c). Class-wise accuracies of the corresponding datasets

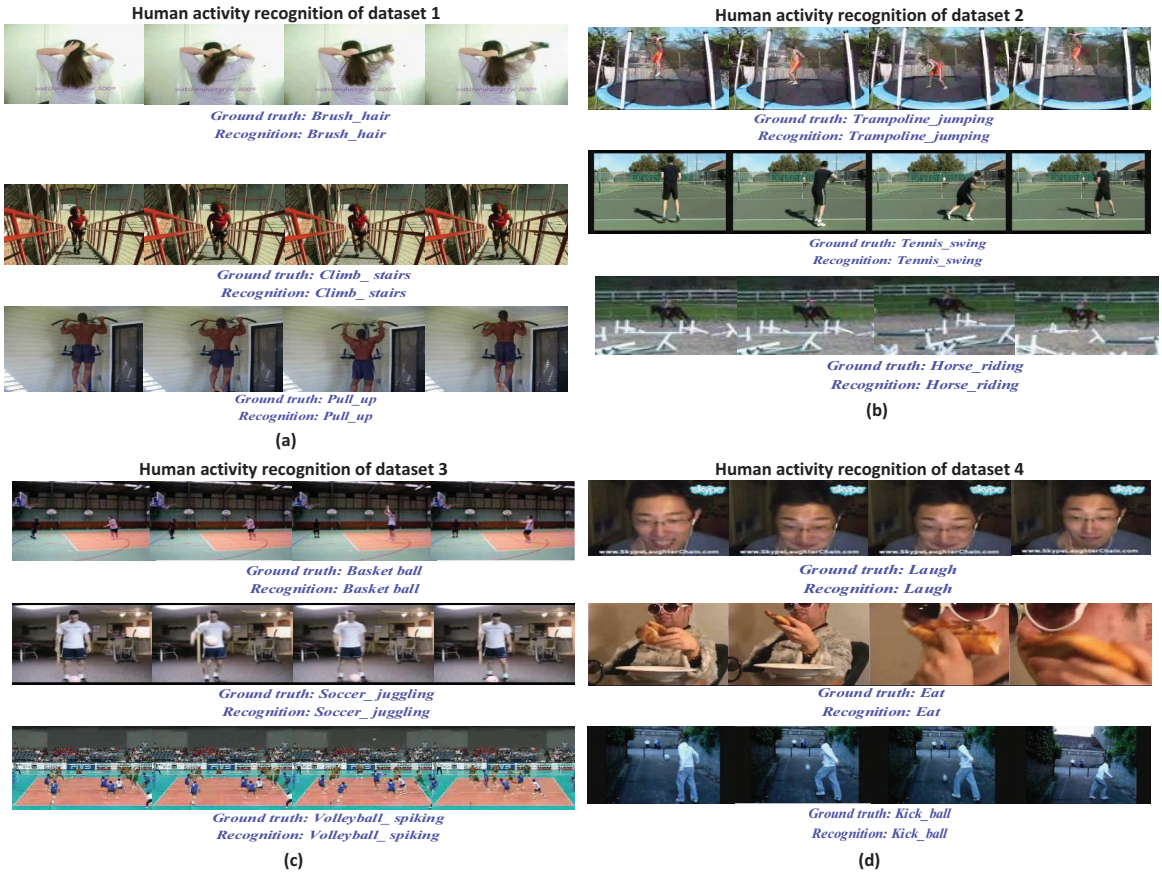

Figure 5. (a), (b), (c) and (d). Activity recognition outcomes using the proposed method

Therefore, the proposed method correctly recognizes the activity.

The investigation underscores the critical influence of input video data on the proposed method. Through accurate activity recognition, this approach enhances surveillance systems by enabling nuanced analysis of human behavior, improving decision-making and bolstering safety and security measures. As a result of its precise recognition capabilities, the proposed model significantly improves overall surveillance system performance.

4.2. Performance comparison with JHMDB and HMDB51 datasets

Comparative evaluations are performed by contrasting the proposed method with corresponding datasets, employing these performance metrics. This evaluation assesses the performance of the MCCBCN system against traditional approaches such as TD-net [10], MF + MVF [13], ViCo-MoCo [15], and Deep BiLSTM [17].

Table 3 shows the accuracy of different methods on a specific dataset. TD-net gets 79.3% accuracy, proving it works well for predictions. Deep BiLSTM reaches 76.3% accuracy, while ViCo-MoCo hits 78.5%. MF + MVF does better with 87.76% accuracy, showing it predicts more. The new method, MCCBCN, beats them all with 99.34% accuracy, showing it's the best at predicting results for this dataset.

Table 4 shows how well different methods perform, based on their accuracy percentages for a specific dataset. ViCo-MoCo has an accuracy of 71.4%. MF + MVF and Bi-GRU have accuracies of 71.28% and 71.89%, respectively. Also, CNN-RNN reaches 70.8% accuracy. The "MCCBCN" (Proposed) method tops the list with a remarkable 99.45% accuracy.

4.3. Performance analysis of human activity recognition

The efficiency is assessed using several metrics across relevant datasets. Comparative evaluations are performed by contrasting the proposed method with

Table 3. Accuracy analysis comparison of the JHMDB dataset

Techniques	Accuracy (%)
TD-net [10]	79.3
MF + MVF [13]	87.76
ViCo-MoCo [15]	78.5
Deep BiLSTM [17]	76.3
MCCBCN (Proposed)	99.34

Table 4. Accuracy analysis comparison of the HMDB51 dataset

Techniques	Accuracy (%)
CNN-RNN [11]	70.8
MF + MVF [13]	71.28
ViCo-MoCo [15]	71.4
CNN Bi-GRU [16]	71.89
MCCBCN (Proposed)	99.45

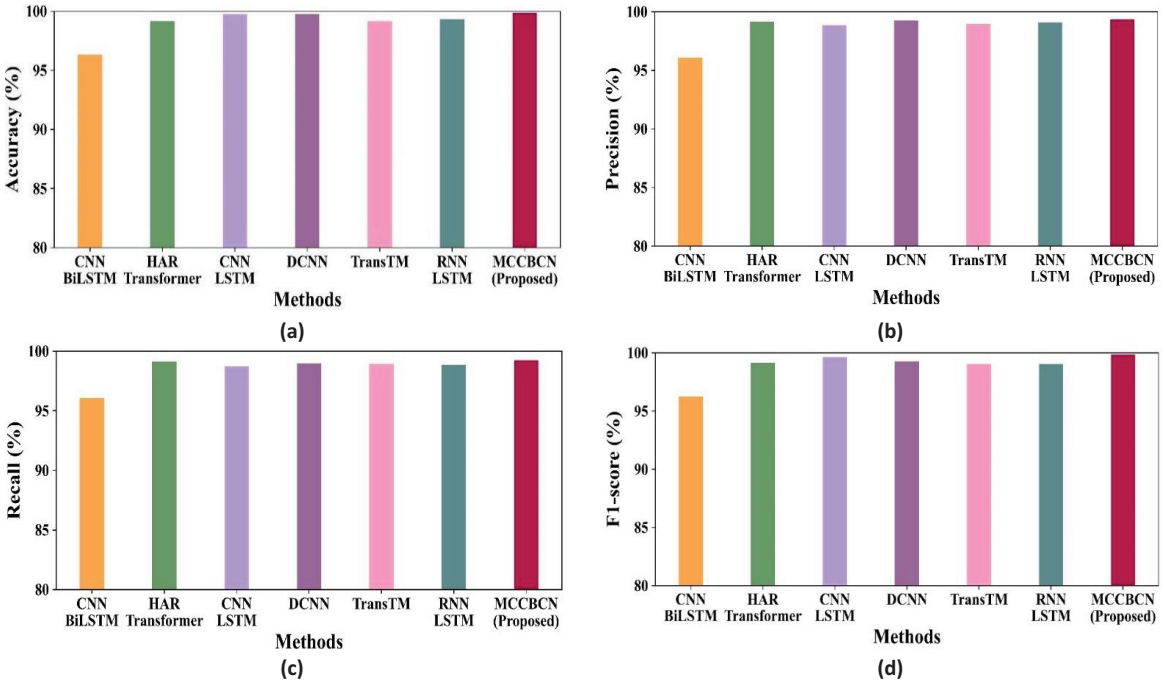

Figure 6. (a) accuracy, (b) precision, (c) recall, and (d) F1-score on recent HAR techniques

established techniques, employing these performance metrics. This evaluation compares the performance of the proposed system with traditional approaches such as CNN BiLSTM [12], HAR transformer [13], CNN LSTM [14], TransTM [18], and RNN LSTM [19].

Figures 6 (a), (b), (c), and (d) illustrate various HAR models assessed using key performance metrics: accuracy, precision, recall, and F1-score. The CNN-BiLSTM model achieved an accuracy of 96.37%, with corresponding precision, recall, and F1-score values of 96.14%, 96.14%, and 96.31%, respectively. The HAR transformer model established high performance across all metrics, attaining an accuracy of 99.2% along with consistent precision, recall, and F1-score values of 99.2%.

Similarly, the CNN-LSTM model showed superior accuracy at 99.8%, with precision, recall, and F1-score values of 98.89%, 98.78%, and 99.7%. The DCNN model achieved an accuracy of 99.80%, with precision, recall, and F1-score values of 99.31%. TransTM achieved an accuracy of 99.2% with a precision, recall, and F1-score of 99%, 99%, and 99.1%, respectively. MCCBCN outperformed with an accuracy of 99.50% and precision, recall, and F1-score values of 99.42%, 99.3%, and 99.9%, respectively.

Fig. 7 illustrates the time complexity of the introduced approach compared to existing techniques like TD-Net, MF+MVF, and ViCo-MoCo. The time complexities for these methods are as follows: TD-Net takes 35 seconds, MF+MVF takes 42 seconds, and ViCo-MoCo takes 27.9 seconds. In contrast, the proposed method, MCCBCN, significantly outperforms these techniques with a time complexity of just 10.2 seconds. The introduced MCCBCN model efficiently surpasses current HAR methods by solving major challenges with contextual perception, temporal dependency modeling, and class imbalance treatment. In contrast to

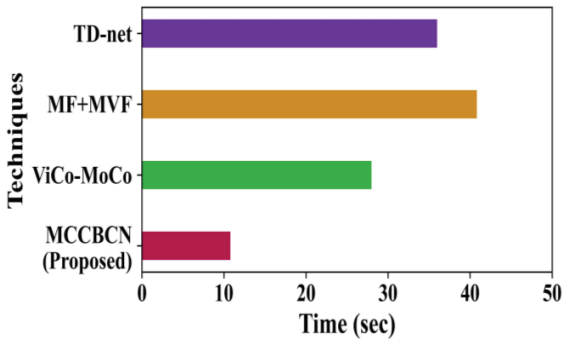

Figure 7. Time complexity analysis

models like CNN-BiLSTM or HAR Transformer, which either do not leverage multi-branch contextual information or are not evaluated across diverse datasets, MCCBCN presents a metapath-based contextual learning approach that improves its capacity to identify intricate and overlapping activities. The incorporation of a bidirectional cascade structure further enhances temporal feature extraction by modeling forward and backward dependencies across activity sequences, a capability not always achieved by most current models. The application of the EFTTOA also simplifies training and enhances learning efficiency with faster convergence and better generalization over diverse datasets.

Additionally, MCCBCN addresses the vital problem of class imbalance by using LSMOST to effectively represent minority-class activities and learn them during training. This ensures the model performs more balanced than other approaches, such as ViCo-MoCo-DL or TransTM, which have not received sufficient validation on diverse or imbalanced datasets. Having an accuracy of 99.50%, an F1-score of 99.9%, and a reasonable time complexity of 10.2 seconds,

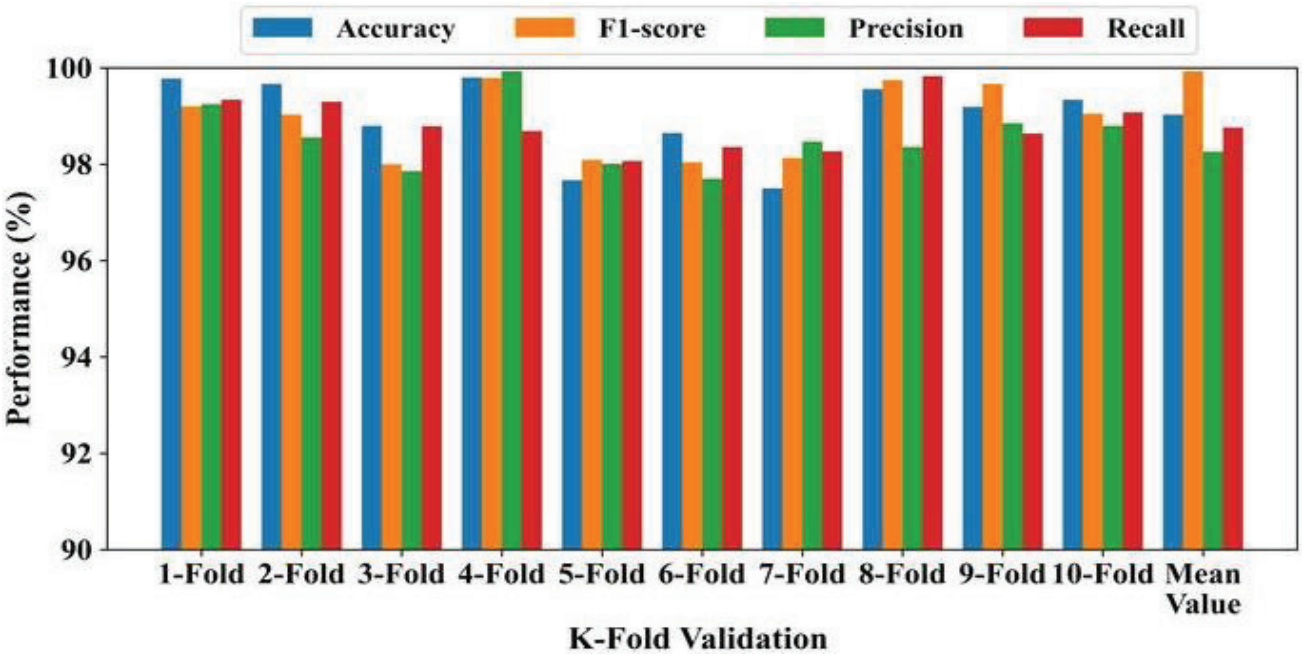

Figure 8. K-fold validation analysis

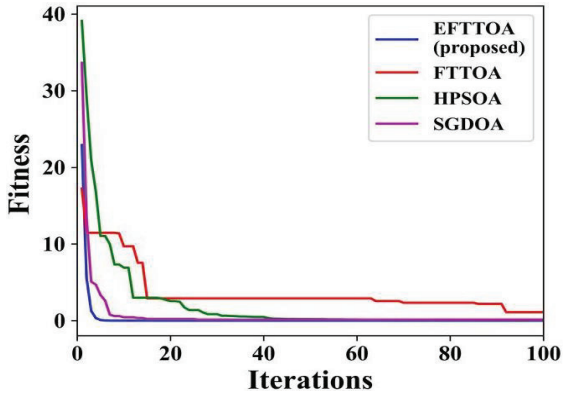

Figure 9. Fitness curve

MCCBCN raises the bar in HAR by achieving robustness, precision, and scalability combined. Its end-to-end design promises higher performance not just in conventional metrics but also in real-world usability, where subtle activity recognition and generalization are paramount.

Fig. 8 depicts performance metrics for the proposed model, MCCBCN, evaluated using the average 10-fold cross-validation.

Fig. 9 compares the performance of four optimization algorithms, like EFTTOA (proposed), FTTOA, Hierarchical Particle Swarm Optimization algorithm (HPSOA) [30], and Stochastic Gradient Descent Optimization Algorithm (SGDOA) [31], over 100 iterations, with fitness values plotted against iterations. The proposed EFTTOA algorithm shows significantly faster convergence, reaching a lower fitness value in fewer iterations, compared to the existing algorithms. This recommends that EFTTOA is more efficient and effective in optimizing the given problem, as it consistently achieves better results more quickly and stabilizes earlier than FTTOA, HPSOA, and SGDOA.

4.4. Ablation study

This research on HAR investigates the impact of removing specific components from the MCCBCN architecture on accuracy, precision, recall, and F1-score across three datasets, such as JHMDB, UCF 11, UCF YouTube action, and HMDB51.

Table 5 shows the performance of the MCCBCN architecture for HAR across four datasets, evaluated alongside various model configurations. The full MCCBCN architecture achieves exceptionally high accuracy across all datasets, ranging from 99.34% to 99.45%, indicating its robustness and effectiveness. When specific components of the MCCBCN architecture are excluded, like metapath context, Bi-directional cascade, EFTTOA Integration, LSMOST, and MobileNetV2-Lite & MGCN Integration, there is a noticeable decrease in accuracy. For instance, taking out the metapath context makes the accuracy fall between 85.9% and 86.7%. This shows how important it is to capture context information to boost performance. In the same way, removing the Bi-directional cascade increases accuracy from 86.2% to 89.3%. This points out how crucial this setup is to make the model more accurate.

5. Conclusion

This research introduced the MCCBCN method, which leverages metapaths to enhance the accuracy and robustness of contextual information capture for HAR. Significant advancements in the field were achieved through the utilization of both convolutional and bi-directional cascade architectures. The model's accuracy and robustness were further boosted by integrating the EFTTOA. Additionally, the LSMOST was employed for dataset expansion, ensuring a comprehensive and balanced representation of minority classes, mitigating biases from imbalanced class distributions, and improving model

Table 5. Ablation study of the introduced method with their datasets

Model Configuration	JHMDB Accuracy (%)	UCF 11 Accuracy (%)	UCF YouTube Action Accuracy (%)	HMDB51 Accuracy (%)
Full MCCBCN architecture	99.34	99. 39	99.42	99.45
Without Metapath Context	85.9	86.7	85.9	85.3
Without Bi-directional Cascade	89.3	86.2	87.7	88.3
Without EFTTOA Integration	89.1	88.1	88.5	87.2
Without LSMOST	88.5	85.5	86.1	87.8
Without MobileNetV2-Lite & GCN Integration	88.9	88.1	86.7	87.7

generalizability. The integration of the MobileNetV2-Lite backbone with MGCN allowed for the extraction of human-centric features, while multi-view feature computation captured diverse activity patterns. Selection criteria based on relative entropy, mutual information, and the SCC ensured the inclusion of the most relevant features, thereby enhancing model performance and interpretability. The performances were recorded as 99.50% (accuracy), 99.9% (F1-score), 99.42% (precision), and 99.3% (recall), respectively. The accuracy comparison between the JHMDB and HMDB51 datasets showed scores of 99.34% for both datasets. The time complexity of the proposed methods is 10.2 seconds. These validate the dominance of the MCCBCN model in achieving highly accurate, robust, and reliable HAR, outperforming existing methods across key evaluation metrics such as FDR and FOR, especially under varying prevalence conditions.

Future scope In future work, we aim to enable the proposed HAR system to support real-time processing, allowing instantaneous analysis and decision-making on widely varying, continuous data streams. The goals will be to modify the MCCBCN architecture for low-latency applications, to combine edge computing so that localized processing takes place, and to investigate novel compression and transmission techniques to efficiently manage big data from sensors. Additionally, the model's flexibility to varying implementations in real-world situations and hardware platforms is ensured to ensure wider use and deployment.

Data availability statements Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHORS

Praveen S. Banasode* – Department of Master of Computer Applications, Jain College of Engineering, Belagavi, Karnataka 590014, India, e-mail: praveenb.jce@gmail.com.

Sunita Padmannavar – Department of Master of Computer Applications, Gogte Institute of Technology, Belagavi 590006, Karnataka, e-mail: sunitapdm@gmail.com.

*Corresponding author

ACKNOWLEDGEMENTS

None

References

[1] Hussain, S. U. Khan, N. Khan, M. Shabaz, and S. W. Baik, "AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107218, Jan. 2024. doi: 10.1016/j.engappai.2023.107218

[2] Kabiruzzaman, M. Shidujaman, S. Hassan Shifat, P. Debnath, and S. Hossain, "Time series analysis of Care Records data for nurse activity recognition in the wild," *Human Activity and Behavior Analysis*, pp. 405–415, Mar. 2024. doi: 10.1201/9781003371540-28

[3] A. Kushwaha, A. Khare, and O. Prakash, "Human activity recognition algorithm in video sequences based on the fusion of multiple features for realistic and multi-view environment," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 22727–22748, Aug. 2023. doi: 10.1007/s11042-023-16364-z

[4] S. Ankalaki and M. N. Thippeswamy, "Optimized convolutional neural network using hierarchical particle swarm optimization for sensor based human activity recognition," *SN Computer Science*, vol. 5, no. 5, Apr. 2024. doi: 10.1007/s42979-024-02794-5

[5] Md. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Multi-level feature Fusion for multimodal human activity recognition in internet of healthcare things," *Information Fusion*, vol. 94, pp. 17–31, Jun. 2023. doi: 10.1016/j.inffus.2023.01.015

[6] T. R. Mim *et al.*, "Gru-Inc: An inception-attention based approach using GRU for human activity recognition," *Expert Systems with Applications*, vol. 216, p. 119419, Apr. 2023. doi: 10.1016/j.eswa.2022.119419

[7] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, "Adventures in data analysis: A systematic review of deep learning techniques for pattern recognition in cyber-physical-social systems," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 22909–22973, Aug. 2023. doi: 10.1007/s11042-023-16382-x

[8] A. Ray, M. H. Kolekar, R. Balasubramanian, and A. Hafiane, "Transfer learning enhanced vision-based human activity recognition: A decade-long analysis," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100142, Apr. 2023. doi: 10.1016/j.jjimei.2022.100142

[9] M. K. Jannat, Md. S. Islam, S.-H. Yang, and H. Liu, "Efficient Wi-Fi-based human activity recognition using adaptive antenna elimination," *IEEE Access*, vol. 11, pp. 105440–105454, 2023. doi: 10.1109/access.2023.3320069

[10] T. Nguyen, D. Pham, H. Vu, and T. Le, "A robust and efficient method for skeleton-based Human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, vol. 16, no. 8, pp. 709–726, Jul. 2022. doi: 10.1049/cvi2.12119

[11] C. Zhang, Y. Xu, Z. Xu, J. Huang, and J. Lu, "Hybrid handcrafted and learned feature framework for Human Action Recognition," *Applied Intelligence*, vol. 52, no. 11, pp. 12771–12787, Feb. 2022. doi: 10.1007/s10489-021-03068-w

[12] S. K. Challa, A. Kumar, and V. B. Semwal, "A multi-branch CNN-BILSTM model for human activity recognition using wearable sensor data," *The Visual Computer*, vol. 38, no. 12, pp. 4095–4109, Aug. 2021. doi: 10.1007/s00371-021-02283-3

[13] I. Dirgová Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable sensor-based human activity recognition with Transformer model," *Sensors*, vol. 22, no. 5, p. 1911, Mar. 2022. doi: 10.3390/s22051911

[14] Mst. A. Khatun *et al.*, "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–16, 2022. doi: 10.1109/jtehm.2022.3177710

[15] T. Shanableh, "Vico-Moco-DL: Video coding and motion compensation solutions for human activity recognition using deep learning," *IEEE Access*, vol. 11, pp. 73971–73981, 2023. doi: 10.1109/access.2023.3296252

[16] T. Ahmad *et al.*, "Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection," *IEEE Access*, vol. 11, pp. 33148–33159, 2023. doi: 10.1109/access.2023.3263155

[17] N. Hassan, A. S. Miah, and J. Shin, "A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition," *Applied Sciences*, vol. 14, no. 2, p. 603, Jan. 2024. doi: 10.3390/app14020603

[18] Y. Liu *et al.*, "TransTM: A device-free method based on time-streaming multiscale transformer for human activity recognition," *Defence Technology*, vol. 32, pp. 619–628, Feb. 2024. doi: 10.1016/j.dt.2023.02.021

[19] A. C. Cob-Parro, C. Losada-Gutiérrez, M. Marrón-Romera, A. Gardel-Vicente, and I. Bravo-Muñoz, "A new framework for deep learning video based human action recognition on the edge," *Expert Systems with Applications*, vol. 238, p. 122220, Mar. 2024. doi: 10.1016/j.eswa.2023.122220

[20] https://figshare.com/articles/dataset/JHMDB/19179260

[21] https://www.kaggle.com/datasets/pypiahmad/realistic-action-recognition-ucf50-dataset

[22] https://www.kaggle.com/datasets/pypiahmad/ucf-youtube-action-data-set

[23] https://www.kaggle.com/datasets/easonlll/hmdb51

[24] K. Wang, T. Zhou, M. Luo, X. Li, and Z. Cai, "Generative adversarial minority enlargement—a local linear over-sampling synthetic method," *Expert Systems with Applications*, vol. 237, p. 121696, Mar. 2024. doi: 10.1016/j.eswa.2023.121696

[25] L. Si *et al.*, "A novel coal-gangue recognition method for top coal caving face based on IALO-VMD and improved mobilenetv2 network," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023. doi: 10.1109/tim.2023.3316250

[26] Z. Chen *et al.*, "Learnable graph convolutional network and feature fusion for multi-view learning," *Information Fusion*, vol. 95, pp. 109–119, Jul. 2023. doi: 10.1016/j.inffus.2023.02.013

[27] X. Fu and I. King, "MECCH: Metapath context convolution-based heterogeneous graph neural networks," *Neural Networks*, vol. 170, pp. 266–275, Feb. 2024. doi: 10.1016/j.neunet.2023.11.030

[28] J. Ye, Z. Yu, Y. Wang, D. Lu, and H. Zhou, "PlantBiCNet: A new paradigm in plant science with bi-directional cascade neural network for detection and counting," *Engineering Applications of Artificial Intelligence*, vol. 130, p. 107704, Apr. 2024. doi: 10.1016/j.engappai.2023.107704

[29] Z. Tian and M. Gai, "Football team training algorithm: A novel sport-inspired meta-heuristic optimization algorithm for global optimization,"

Expert Systems with Applications, vol. 245, p. 123088, Jul. 2024. doi: 10.1016/j.eswa.2023.123088

[30] Z. Li, Y. Liu, B. Liu, J. Le Kernec, and S. Yang, "A holistic human activity recognition optimisation using AI techniques," *IET Radar, Sonar & Navigation*, vol. 18, no. 2, pp. 256–265, Sep. 2023. doi: 10.1049/rsn2.12474

[31] I. Priyadarshini, R. Sharma, D. Bhatt, and M. Al-Numay, "Human activity recognition in cyber-physical systems using optimized machine learning techniques," *Cluster Computing*, vol. 26, no. 4, pp. 2199–2215, Aug. 2022. doi: 10.1007/s10586-022-03662-8