IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 September 2024, accepted 9 October 2024, date of publication 16 October 2024, date of current version 5 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3481631

RESEARCH ARTICLE

Multimodal Human Action Recognition Framework Using an Improved CNNGRU Classifier

MOUAZMA BATOOL[1], MONEERAH ALOTAIBI[2], SULTAN REFA ALOTAIBI[2], DINA ABDULAZIZ ALHAMMADI[3], MUHAMMAD ASIF JAMAL[1,4], AHMAD JALAL[1,5], AND BUMSHIK LEE[4]

[1]Faculty of Computing and AI, Air University, Islamabad 44000, Pakistan
[2]Department of Computer Science, College of Science and Humanities Dawadmi, Shaqra University, Shaqra 11961, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[4]Department of Information and Communication Engineering, Chosun University, Gwangju 61452, Republic of Korea
[5]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, Republic of Korea

Corresponding author: Bumshik Lee (bslee@chosun.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00217471). The author would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work. This work was supported through Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R508), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Activity recognition from multiple sensors is a promising research area with various applications for remote human activity tracking in surveillance systems. Human activity recognition (HAR) aims to identify human actions and assign descriptors using diverse data modalities such as skeleton, RGB, depth, infrared, inertial, audio, Wi-Fi, and radar. This paper introduces a novel HAR system for multi-sensor surveillance, incorporating RGB, RGB-D, and inertial sensors. The process involves framing and segmenting multi-sensor data, reducing noise and inconsistencies through filtration, and extracting novel features, which are then transformed into a matrix. The novel features include dynamic likelihood random field (DLRF), angle along sagittal plane (ASP), Lagregression (LR), and Gammatone cepstral coefficients (GCC), respectively. Additionally, a genetic algorithm is utilized to merge and refine this matrix by eliminating redundant information. The fused data is finally classified with an improved Convolutional Neural Network - Gated Recurrent Unit (CNNGRU) classifier to recognize specific human actions. Experimental evaluation using the leave-one-subject-out (LOSO) cross-validation on Berkeley-MHAD, HWU-USP, UTD-MHAD, NTU-RGB+D60, and NTU-RGB+D120 benchmark datasets demonstrates that the proposed system outperforms existing state-of-the-art techniques with the accuracy of 97.91%, 97.99%, 97.90%, 96.61%, and 95.94% respectively.

INDEX TERMS Convolutional neural network, depth camera, human action recognition, inertial sensors, multi-sensors, RGB.

I. INTRODUCTION

Multi-sensor human activity recognition (HAR) has become popular in surveillance, robotics, security, gaming, and healthcare, among other fields [1], [2], [3], [4], [5]. Accurately capturing human movements and activities has advanced due to the integration of several sensors, such as inertial and vision [6]. But controlling crowded backgrounds, integrating heterogeneous data, correcting occlusions, and handling complex and comparable actions are just a few of the significant obstacles that modern multi-sensor surveillance systems must overcome [7]. The precision and dependability of HAR systems are directly impacted by these difficulties [8], [9], [10].

Researchers have turned to depth cameras to address privacy concerns associated with RGB video cameras in

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

human activity detection [11], [12], [13]. Depth cameras can generate valuable 3D information but rely on infrared light, making them impractical for outdoor use [14], [15], [16], [17]. They are also more expensive and less commonly found in average households. On the other hand, wearable sensors offer continuous monitoring and unlimited observation space, overcoming many limitations of vision-based sensors [18]. With advancements in microelectromechanical (MEMS) systems, wearable sensors can now be integrated into smartphones and smartwatches [19], [20], [21]. However, they still face challenges related to sensor orientation and placement on the human body.

Integrating both sensor modalities enhances recognition accuracy in real-time applications, as vision and inertial sensors provide complementary information [22]. Vision and wearable sensors have been combined in various real-world applications, including elderly assistance, patient monitoring, and exergaming. For HAR in surveillance systems, researchers have frequently combined wearable and vision sensors [23]. These studies have influenced multimodal sensors, including RGB, depth, and inertial sensors, in the HAR model [24], [25], [26], [27].

Handcrafted and deep-learning classifiers have been employed to recognize HAR in surveillance systems [28]. Handcrafted features, requiring expert knowledge and precise calculations, limit the performance of HAR techniques [29]. Deep learning techniques, such as convolutional neural networks (CNNs), artificial neural networks (ANNs), and recurrent neural networks (RNNs), automate feature extraction and efficiently process sequential information [30]. Despite their advantages, deep learning classifiers still face challenges like vanishing and exploding gradients [31], [32], [33], [34], [35].

Machine learning (ML) and deep learning (DL) approaches have shown great promise in recent years for improving HAR, creating a scalable and efficient architecture for multi-sensor systems is still a challenging task. Getting useful features from multi-sensor data is essential for making optimizers and classifiers work better [36], [37]. Nevertheless, creating the best characteristics is difficult and requires sophisticated techniques, especially when dealing with visual and inertial data. The goal of this research is to improve the use of multi-sensor data by creating robust and effective classifiers that open up new application possibilities. Enhancing the theoretical and practical features of HAR in surveillance systems while addressing the major issues brought on by complicated scenarios is the main goal of this doctoral research project [38], [39], [40]. Achieving high accuracy and efficiency is the ultimate goal because these are essential for reliable, real-time computer vision applications in human activity identification.

The current study introduces innovative feature extraction methodologies and a new deep learning classifier for human action recognition, utilizing diverse features from inertial, RGB, and depth images. To this end, five benchmark datasets have been utilized. The results demonstrate the effectiveness of combining multimodal approaches, underscoring their potential to enhance human action recognition performance through deep learning. The main contributions of this research can be summarized as follows:

- The existing state-of-the-art features, including Markov random field (MRF), autoregression (LR), mel-frequency cepstral coefficients (MFCCs), have been enhanced and modified to dynamic likelihood random field (DLRF), lag regression (LR), Gammatone cepstral coefficients (GCC), respectively.
- The genetic algorithm (GA) is utilized to fuse data from multimodal sensors to recognize complex human action patterns, offering both contextual information and behavior classification.
- Finally, the novel Convolutional Neural Network - Gated Recurrent Unit (CNNGRU) classifier, an enhanced version of the convolutional neural network (CNN), has been employed to categorize the Berkeley-MHAD, UTD-MHAD, HWU-USP, NTU-RGB+D60, and NTU-RGB+D120 benchmark datasets, achieving significantly superior results compared to other state-of-the-art methods.

The subsequent sections of the paper are organized as follows: Section II provides a comprehensive overview of the latest frameworks used in human action recognition. Section III introduces the modified feature extraction techniques designed to highlight the most significant aspects of multimodal data. Section IV presents the experimental findings, details the datasets used, and compares the results with current state-of-the-art methods. Finally, Section V summarizes the research and outlines potential directions for future investigation.

II. RELATED WORKS

Computer vision techniques for human action recognition using inertial, RGB, and depth sensors have encountered several challenges, such as sensitivity to orientation, background clutter, and limited field of view. To address these issues, researchers have increasingly integrated multiple sensing modalities into various smart devices, including the Apple Watch and OpenPose. This section reviews the literature on machine learning and deep learning approaches for robust human action recognition, highlighting methods that utilize data from diverse sensors.

Yang et al. [41] proposed a multimodal learning model for human action recognition that combines skeleton data with depth sequential information entropy maps (DSIEM) through action fusion. Evaluated on the UTD-MHAD dataset, the model achieved an accuracy of 88.37%. However, the approach might lack interpretability and be ineffective in recognizing activities outside the trained model. Golestani et al. [42] evaluated and compared six different machine learning classifiers, finding that the random forest classifier achieved an average accuracy of 91.5% on the Berkeley MHAD dataset. Nevertheless, testing across multiple datasets is essential to capture a range of simple and complex activities

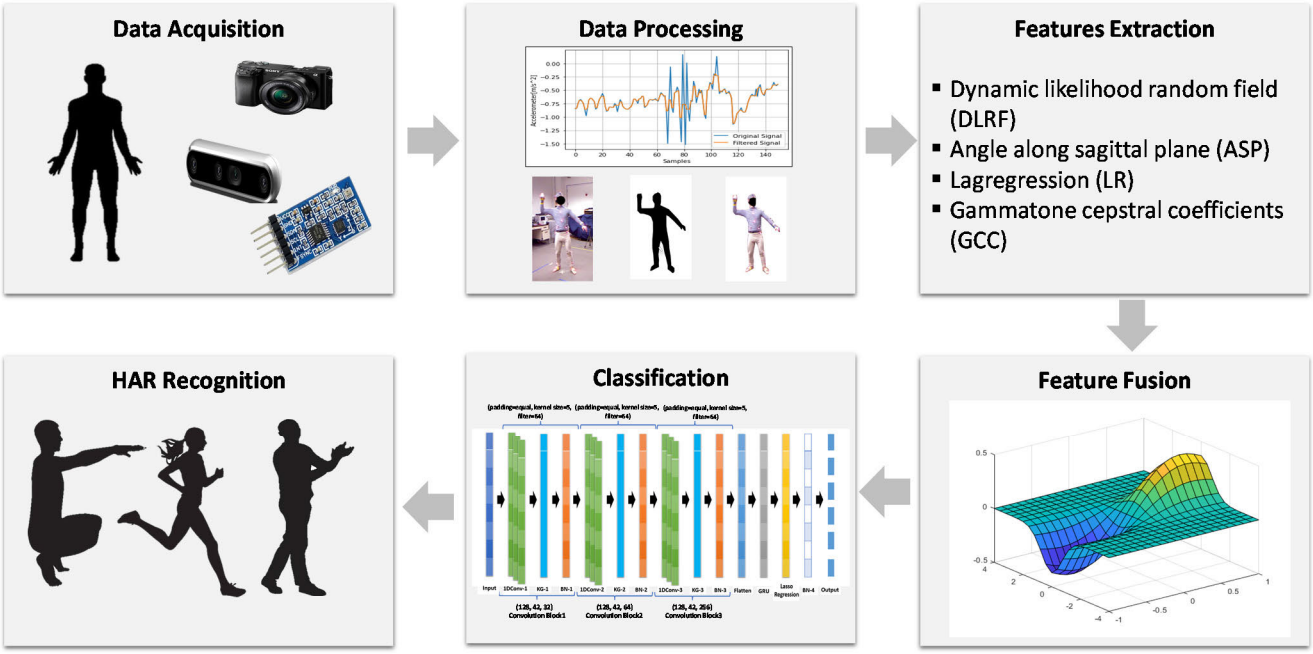


FIGURE 1. Graphical representation of the proposed model utilizing multimodal sensors.

for a thorough evaluation. Javeed et al. [43] presented an innovative multimodal IoT-based method for classifying locomotion using the HWU-USP and Opportunity++ datasets. This approach employs a recursive neural network that integrates data from inertial, ambient, and vision-based sensors. Using a 10-fold cross-validation technique, the method achieved an impressive accuracy of 87%. Kang et al. [44] employed a CNN for accelerometer data and an LSTM-RNN for RGB data, integrating the outputs with a deep fusion network to achieve a 93% accuracy on the Berkeley MHAD dataset. Despite its high overall performance, the model showed suboptimal results in certain activities, likely due to the vanishing gradient problem within the proposed architecture. Ni et al. [45] proposed a skeleton-to-sensor knowledge distillation (PSKD) model that integrates the human skeleton and time-series data. Although the model achieved accuracies of 94.76% on the Berkeley MHAD dataset and 95.19% on the UTD-MHAD dataset, it faces challenges regarding computational complexity and limited interpretability. Ranieri et al. [46] showcased a deep learning (DL) framework that integrates inertial, visual, and ambient sensors. When tested on the HWU-USP and UTD-MHAD datasets, the model achieved accuracies of 93.75% and 92.33%, respectively. However, its sensitivity to parameter settings require careful tuning to achieve optimal performance. William et al. [47] introduced a model utilizing accelerometer and gyroscope sensors, which achieved a 95% accuracy on the UCI HAR dataset using seven machine learning classifiers. However, its practical applicability might be limited due to the substantial computational resources required for both simple and complex scenarios. Khan et al. [48] developed a human activity recognition model using accelerometer data, incorporating ensemble bagged trees, K-Nearest Neighbor (KNN) [49], and Support Vector Machine (SVM) [50] classifiers. However, the SVM component demonstrates a notable sensitivity to parameter configurations, underscoring the critical importance of meticulous tuning. While the model exhibits superior performance in specific activity recognition tasks, its capacity for generalization across a broad spectrum of daily activities occurring in diverse environmental contexts and exhibiting variable patterns may be constrained. This limitation potentially impacts the robustness of the model and adaptability in real-world applications where activity diversity and environmental variability are prevalent.

III. PROPOSED MODEL

Multimodal framework for human action recognition using inertial, RGB, and depth sensors is proposed. The structure of the proposed system includes several stages: preprocessing, feature extraction, feature fusion, and classification, utilizing a GA and a modified CNNGRU classifier. The model is tested and trained on the five benchmark datasets using the Leave-One-Subject-Out (LOSO) [51] validation method. Figure 2 illustrates our proposed framework.

A. DATA PROCESSING

The multimodal sensor data underwent a rigorous data preprocessing stage. Extraction of silhouettes from the RGB and depth image data is performed by applying Laplacian fitting, as described in [51], [52], and [53]. Concurrently, the inertial data underwent preprocessing utilizing a Kaiser window

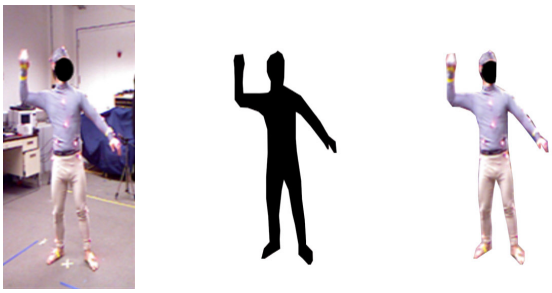
low-pass filter, following the methodology outlined by [54], [55], and [56]. A detailed exposition of each preprocessing technique is provided in Section 2.3, elucidating the specific parameters and implementation procedures employed in this study.

1) RGB-D SILHOUETTE EXTRACTION VIA LAPLACIAN FITTING

The silhouette is abstracted from RGB and depth images by using Laplacian fitting approach [57]. The process starts with the computation of Stauffer's background subtraction using a Gaussian mixture [58], [59], [60]. Next, Laplacian matrices are calculated, and minimized to extract Eigen mattes coefficients. Furthermore, RANSAC algorithm [61] is applied to fit the Eigen mates with inliers of the estimated blob. Finally, optimized matte fitting is used to extract the human silhouette as (1).

$$arg_x min \, \|\varepsilon x - \varphi\|^2 \tag{1}$$

where x is the weight vector, ε is eigenvector, φ presents initial motion estimation. Figure 2 shows the final output.



(a) (b) (c)

FIGURE 2. The output of silhouette extraction over RGB data. (a) original image, (b) extracted silhouette, (c) extracted silhouette from RGB image.

2) INERTIAL DATA PREPROCESSING VIA KAISER WINDOW LOW PASS (LP) FILTER

The Kaiser windowed LP filter [47] was chosen for inertial data preprocessing as it could control the stop band, passband, and transition band with a specific adjustment of filter order. Furthermore, the properties of having a unique sin shape function give the advantages of low side lobes and bandwidth at equal filter length compared to other filters [62], [63]. The output of The Kaiser windowed LP filter is obtained using (2).

$$b(n) = \frac{B_0\left(\frac{2S}{O_f}\sqrt{n(O_f - 1)}\right)}{B_0(S)} \tag{2}$$

where O_f is the order of the filter, B_0 is the Bessel function of zero order, and S is the shape of the window, which is a non-negative real number. Figure 3 shows the output of the filter.

The original signal (blue line) in the accompanying figure displays noticeable noise and fluctuations. In contrast, the filtered signal (orange line) is considerably smoother with fewer spikes. The filtering process has effectively removed noise and outliers, stabilizing the signal and simplifying analysis. Specifically, the extreme peaks and troughs in the original signal have been significantly reduced in the filtered signal, resulting in a more consistent and less irregular waveform.

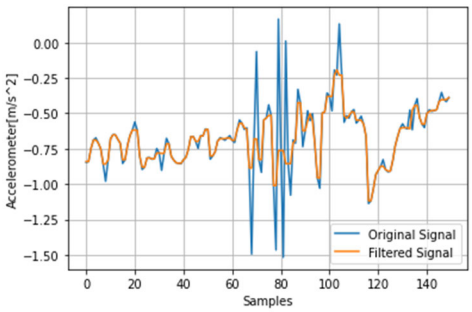


FIGURE 3. The output of Kaiser windowed LP filter over inertial sensors data. The red and blue signal data indicate the filtered and unfiltered data.

B. FEATURE EXTRACTION

The proposed system introduces four novel feature extraction techniques for RGB-D and inertial sensor data. These techniques include dynamic likelihood random field (DLRF) and angle representation along the sagittal plane (ASP). We provide a detailed description of these novel features below. Our analysis indicates that these features demonstrate superior proficiency compared to existing state-of-the-art approaches.

1) RGB-D: DYNAMIC LIKELIHOOD RANDOM FIELD (DLRF)

The dynamic likelihood random field (DLRF) is the modified version of MRF. The MRF calculates the same color pixels into one region [64]. Furthermore, the MRF also integrates over-segmented regions to suit adjusted regions [65], [66], [67], [68]. However, the performance of MRF is affected by new incoming values. Hence, we propose a novel DLRF as an integral component of our architecture in this study. This innovative approach employs simulated annealing instead of traditional Gibbs sampling techniques. The primary objectives of this methodological shift are twofold: to enhance the efficiency of chain mixing and to mitigate the correlation between consecutive samples [69], [70]. These improvements are designed to optimize the overall performance and reliability of our proposed system. The proposed DLRF algorithm is described in Algorithm 1, and Figure 4 shows the results of the proposed DLRF.

The input to DLRF is represented as $\boldsymbol{G_d}$. The DLRF first selects the given image's pixel $\boldsymbol{E_c}$ at random and searches all nearby interconnected pixels $\boldsymbol{y}$, represented by $\boldsymbol{Q}(\boldsymbol{E})$. Next, the probability distribution and energy function of every pixel are computed for the set of individual pixels $\boldsymbol{E_u} = \{\boldsymbol{e_i}\}_{\boldsymbol{u}}$ and the pair of pixels grouped together $\boldsymbol{E_w} = \{\boldsymbol{e_i}, \boldsymbol{e_j}\}_{\boldsymbol{v}}$, which is

Algorithm 1 Dynamic Likelihood Random Field (DLRF)

Input: Initial parameter setting

do

$$Q(E) = \sum_{d\epsilon y} \prod_{d=1}^{n} G_d\,(E_c)$$

$$F(E) = \frac{1}{Q(E)} \exp^{\left(\sum_{j=1}^{V} V(E_u) + \sum_{j=1}^{W} W(E_W)\right)\omega t}()$$

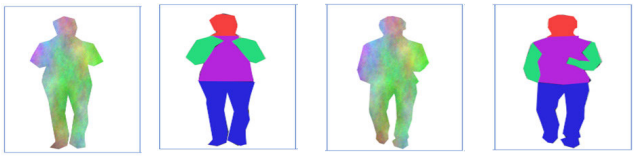
$$Q\,(E_v) = \frac{1}{Q(E)} Prob\,(e_1)\,Prob\,(e_2)\,Prob\,(e_3)\ldots$$

$$Q\,(E_w) = \frac{1}{Q(E)} Prob\,(e_1, e_{new}, T)\,,\ t_0 \le T \le t_f$$

$$Q(E) = -log\left[\sum_{j=1}^{V} V\,(E_v) + \sum_{j=1}^{w} W\,(E_w)\right]$$

While (condition not satisfied)

return: final estimated parameters prediction



(a) (b) (c) (d)

FIGURE 4. The MRF results of (a) & (c) and DLRF results of (b) and (d) on arm curl and jogging activities.

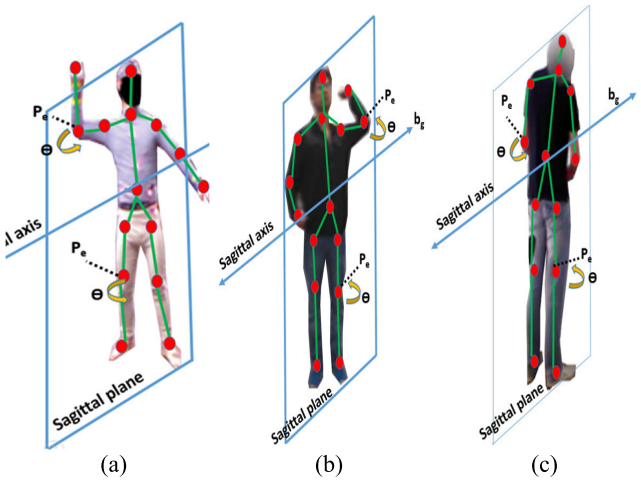
depicted by $F\,(\boldsymbol{E})$. For approximating probability $\boldsymbol{Prob}\,(\boldsymbol{e_n})$ of n pixel $\boldsymbol{e}$, Gibbs sampling is a good inference method. Up till the log-likelihood of the train data is minimized, the process is repeated. The simulated annealing algorithm, on the other hand, is based on the DLRF, which is the concept of a temperature variable that grows from time $\boldsymbol{t_0}$ until the final temperature $\boldsymbol{t_f}$ condition is met for initially pixel $\boldsymbol{e_1}$, new pixel $\boldsymbol{e_{new}}$ at time $\boldsymbol{T}$. Only if the updated results outperform the current ones will the solution be modified.

2) RGB-D: ANGLE ALONG SAGITTAL PLANE (ASP)

By capturing the angle along the sagittal plane, the individual could get several advantages, including a full understanding of a body's 3D motion, which is critical for assessing overall performance, identifying compensatory movements, and creating more effective training or rehabilitation programs. Measuring joint angles in all three planes allows for early detection of abnormal movement patterns that can lead to injury [71], [72], [73]. Moreover, individuals can benefit from personalized training programs that target specific joint angles in different planes. Measuring joint angles in multiple planes can enhance diagnostic accuracy for conditions like scoliosis, joint deformities, or musculoskeletal disorders. The sagittal plane divides the body virtually into left and right halves along the midsagittal plane at the hip, knee, and ankle positions [74]. The angle is measured from the standing posture to its doing activity position and is determined as follows:

$$q = n \cdot \left(\frac{d\theta}{du}\right), n = b_g \times p_e \tag{3}$$

where $\boldsymbol{b_g}$ is the angle formation of bones along the sagittal plane and $\boldsymbol{p_e}$ is the orthogonal distance across the joint position, and the output is depicted in Figure 5.



(a) (b) (c)

FIGURE 5. Measurement of sagittal angles in (a), (b), and (c) at knee, and ankle joints in the midsagittal plane.

3) INERTIAL: LAGREGRESSION (LR)

The lagregression (LR) is the modification of autoregression (AR) [75] feature that incorporates least square approximation and Lagrange multiplier to overcome the limitations of AR. The time-series based AR model depicts each signal sample as a linear combination of past symbols [2], [76], [77]. The AR model functions using static signal data, rendering it ineffective for analyzing real-time series data. The LR can be obtained as (4).

$$LR(c) = \sum_{j=1}^{m} d_j c_{j-1}\,(1-b)^2 - \lambda \tag{4}$$

where c and d are input data and coefficients parameter, respectively, λ is a lagrange multiplier of the order m parameter. The output is depicted in Figure 6.

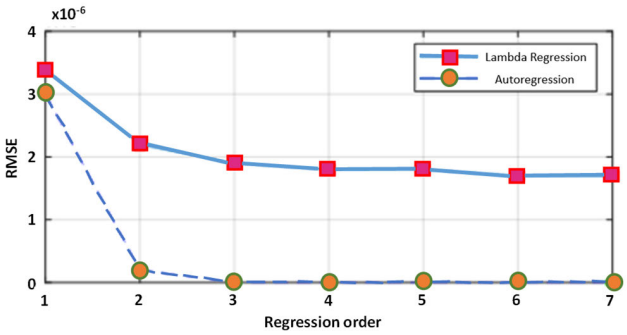


FIGURE 6. The output results of the autoregression and modified Lagrange (LR) regression features.

4) INERTIAL: GAMMATONE CEPSTRAL COEFFICIENTS (GCC)

The mel frequency cepstral coefficients (MFCCs are frequently employed in speech processing to analyze

low-frequency speech resolution [78]. However, having relatively similar energy distribution in both inertial and speech signals makes MFCC useful in the HAR features extraction process. Nonetheless, it demonstrates susceptibility to noise and presents constraints in capturing the dynamic fluctuations in inertial sensor data [79], [80], [81].

The MFCC has been modified to GCC. In GCC the cubic operations are used in place of log operations, and a gammaton filter is used in place of the triangle filter. First, GCC amplifies the high-frequency signal data. Afterwards, the signal undergoes segmentation into overlapping frames, which is followed by the application of fast Fourier transfor (FFT) to acquire the frequency spectrum [82], [83]. Subsequently, the mel filter bank f_{bank} is applied, which is comprised of triangular filters utilized on the frequency spectrum. Finally, the discrete cosine transform (DCT) is employed to compute the MFCC coefficients based on the logarithm of the filterbank energies f_{log}. The details have been specified in Algorithm 2.

Algorithm 2 The MFCC Algorithm

Input: Initial parameter setting
do
$$\mathrm{E} = FFT\left(v\left(a_i\right) \cdot a_i\right)$$
$$f_{log} = log\left(f_{bank} {}^{*} F_{spec}\right)$$
$$MFCC = DCT\left(f_{log}\right)$$
While (condition not satisfied)

return: final predicted coefficients of MFCC

The proposed GCC introduces the Gammatone filter and cubic operations over the triangular filter and log operations, respectively. Algorithm 3 shows the proposed GCC algorithm.

Algorithm 3 The GCC Algorithm

Input: Initial parameter setting
do
$$\mathrm{E} = FFT\left(v\left(a_i\right) \cdot a_i\right)$$
$$Q = \left(c^{n-1}\right)\left[exp\left(2\pi cf_p\right)\right]\left[cos\left(2\pi af_c + P\right)\right] \cdot E$$
$$GCC = DCT(Q)$$
While (condition not satisfied)

return: final predicted coefficients of GCC

In the GCC, cubic operations are used in place of log operations, and a gammatone filter is used in place of the triangle filter [84]. a_i is the input data. The parameter c denotes the variable, f_p signifies the bandwidth parameter, f_c indicates the center frequency, and P stands for the number of phases [85]. The GCC outperforms the cutting-edge MFCC method due to its flexible frequency scaling across the frequency spectrum with limited frequency resolution [86], [87], [88]. Moreover, the cubic operation efficiently manages complex transformations as a non-linear function, as shown in Figure 7. Figure 7(a) displays the MFCC amplitude variations of the signal over time, with distinct peaks and troughs indicating changes in the signa's intensity. Figure 7(c) shows the MFCC variations of the signal's frequency components over time. Figure 7(e) depicts the rate of change of spectral features over time on the Mel scale, with the x-axis representing time and the y-axis representing frequency. Figure 7(b) represents gammatone filter bank of GCC. Figure 7(d) depicts the GCC variations of the signal's frequency components over time. Figure 7(f) depicts the GCC rate of change of spectral features over time, with the x-axis representing time and the zy-axis representing frequency.

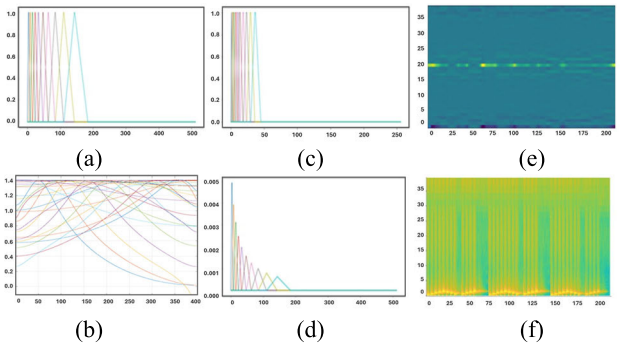


FIGURE 7. The coefficient results for standard MFCC and the modified MFCC. (a) Mel filter bank filter result of MFCC, (b) Filter bank result of GCC, (c) Log energy result of MFCC, (d) Log energy result of GCC, (e) normalized coefficient results of MFCC, & (f) normalized coefficient results of GCC.

C. GENETIC ALGORITHM–FEATURES FUSION

GA [89] has been used as a feature fusio and generally uses three methods, including crossover, iterative selection, and mutation, to drive the evolution of the new population. With each new generation, a fresh set of chromosomes is generated, incorporating the strongest genes from the preceding generation to yield an improved solution. Soltana et al. [90] proposed a novel coding strategy for the simultaneous selection of features and optimal fusion scheme. This coding strategy divides the chromosome into Part A and Part B. With N features, Part A comprises N gene positions, each corresponding to a feature, represented by integer values: 1 indicates an active feature used in feature-level fusion; 0 indicates an active feature used in score-level fusio; and -1 indicates an inactive feature. Part B encodes the fusion model, which depends on the number NF of active features at the feature-level fusion. The fusion strategy and the method by which these extracted features are combined have been outlined in Algorithm 4.

The results of the GA for feature fusion have been depicted in Figure 8. The gradual changes in Figure 8 of the surface plot indicate how smoothly the sensor readings changed across different input data. Figure 8 illustrates that feature fusion has effectively adjusted the feature values, enhancing the ability to capture the underlying patterns in the data [91], [92]. As each iteration progresses, the data representation improves, transitioning from blue to yellow in the diagram.

Algorithm 4 The Overview of Genetic Features Fusion Algorithm

Input: Multimodal Features of RGB, Depth, and Inertial Data do

Initialization: Initialized population of chromosomes with multimodal features

Evaluation: evaluate selected features and calculate fitness score

Selection: Select chromosomes based on fitness scores

Crossover: Select chromosomes based on fitness scores

Mutation: Randomly alter features for variability

Iteration: randomly alter features for variability

While (repeat process until convergence)

return: Final fused features with highest fitness score (optimal fused features)

This process reduces the loss and increases the discriminative power of the features.

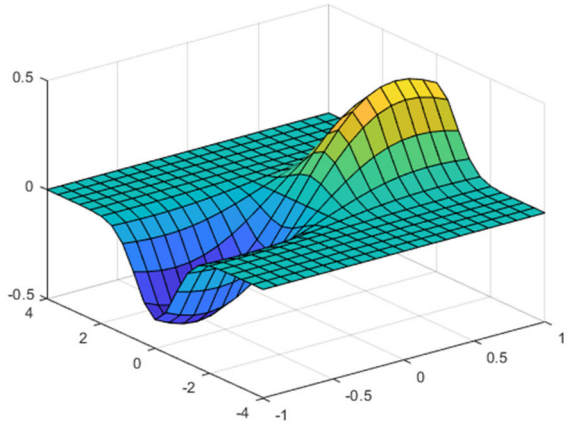


FIGURE 8. The fusion results of GA on inertial, RGB, and depth sensor data.

D. IMPROVED CNNGRU CLASSIFIER

CN [93] stands as a prominent deep learning algorithm primarily employed for image classification tasks [94], [95], [96]. Comprising convolutional, pooling, activation, dropout, and fully connected layers, CNN operates by employing convolution to extract pertinent features. While CNN offers the advantage of concurrently classifying various complex features, it may encounter the issue of dying Rectified Linear Unit (ReLU) [97]. In dying ReLU, the learning process halts if neurons receive negative input during training. To address this, techniques like Leaky ReLU or Parametric ReLU are utilized, permitting small negative inputs to traverse the training model [98], [99], [100]. Nonetheless, Leaky ReLU tends to linearize the output gradually with increased negative input, necessitating manual parameter adjustment to fine-tune its hyperparameters.

In the proposed model, a novel CNNGRU classifier is employed, utilizing Kalman gain (KG) instead of the ReLU layer to forecast the subsequent value, thus addressing the limitations associated with ReLU, leaky ReLU, and parametric ReLU. Unlike ReLU, which forwards input only if it exceeds zero, KG, represented by KG-1 to KG-3, evaluates the incoming value's weight concerning the previous value [101]. Batch normalizatio (BN) layers is represented by BN-1 to BN-4. The D_n signifies the predicted covariance matrix, I denotes the measurement matrix, and I^U is the transpose of the measurement matrix. Consequently, instead of simply passing positive values to the next layer, the KG assesses the significance of the current value based on its certainty relative to the previous value, selectively passing it to the next state or discarding it [102], [103]. Furthermore, the output from the modified CNN is directed to a GRU layer, subsequently feeding into a lasso regression layer for effective data classification [104], [105]. Here, n denotes the number of training instances, p^j signifies the predicted output, x^j represents the actual output, w^j denotes the weighted coefficients, and λ signifies the regularization parameter. Detailed specifications of the novel CNNGRU layers and parameters are provided in Table 1. Figure 9 and Figure 10 show the structure and outputs of the proposed CNNGRU, respectively. Figure 9 shows the input data has dimensions of (128, 42, 32). The network architecture includes three convolutional layers, each followed by BN and KG steps. These convolutional layers extract features from the input data while progressively reducing its spatial dimensions, which helps lower computational costs and mitigate overfitting. A GRU with Lasso regression and batch normalization (BN-4) layer is also incorporated to aggregate the features extracted by the convolutional layers and produce the final output through the final layer.

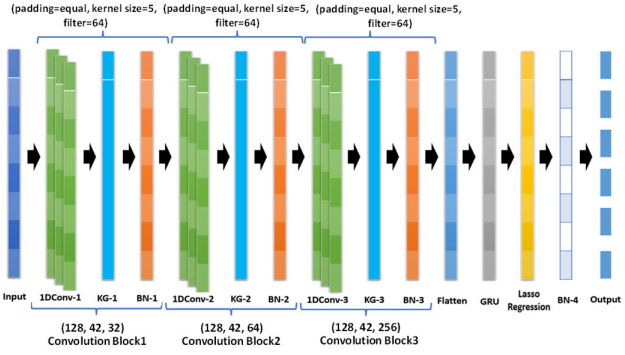


FIGURE 9. Graphical representation of the novel CNNGRU classifier.

IV. EXPERIMENTAL SETUP AND RESULTS

This section provides detailed information on the datasets, experimental findings, accuracy in recognition, and a comparison between our approach and current cutting-edge human activity recognition systems.

TABLE 1. The detailed parameters of the proposed CNNGRU classifier.

Parameters of the proposed CNNGRU classifier	
Layers	Attributes
1DConv-1	padding=equal, kernel size=5, filter=64
KG-1	$KG1 = D_n \cdot I^U \div (I \cdot D_n \cdot I^U + N_{cm})$
BN-1	Parameters set by default values
Conv 1D_2	filter=128, kernel size=5, padding=equal
KG-2	$KG2 = D_n \cdot I^U / (I \cdot D_n \cdot I^U + N_{cm})$
BN-2	Parameters set by default values
Conv 1D_3	padding=equal, kernel size=5, filter=256
KG-3	$KG3 = D_n \cdot I^U / (I \cdot D_n \cdot I^U + N_{cm})$
BN-3	Parameters set by default values
Flatten_1	Flatten layer
GRU	64 units, recurrent activation=sigmoid
Lasso Regression	$LR = n \sum_{j=1}^{n} (p^j + x^j)^2 + \lambda \sum_{j=1}^{m} w_j$
BN-4	Parameters set by default values
Output	Softmax activation layer

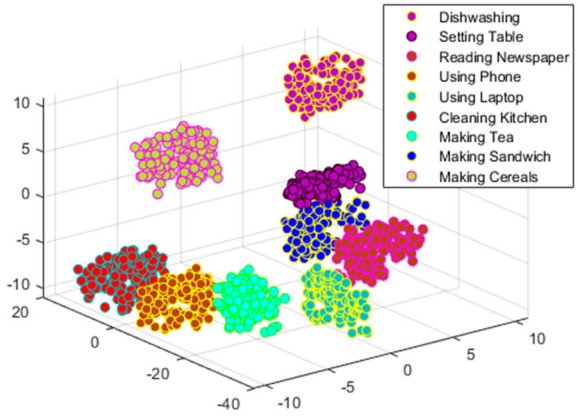


FIGURE 10. Classification outcomes of inertial, RGB, and depth data on the HWU-USP dataset.

A. DATASET DESCRIPTION

1) THE UTD-MHAD DATASET

Chen et al. [49] implemented the UTD-MHAD dataset at the University of Texas Dallas. The dataset comprises 27 distinct actions performed by 8 participants, evenly divided between 4 males and 4 females, within an indoor setting. Each participant executed each action four times, totaling 861 sequences. RGB, skeleton, inertial, and depth data were recorded using wearable and Kinect sensors. The actions include swiping left (SL) and right (SR) with the right arm, drawing clockwise (DC) and counterclockwise (DA) circles with the right hand, sitting (SI), standing (ST), waving with the right hand (WR), clapping (CL), bowling (BW), boxing (BX), pushing (PU), catching (CT), throwing with the right arm (TR), jogging (JG), squatting (SQ), walking (WK), throwing (TH) or picking up objects with the right hand (PR), forward lunging (FL), knocking on a door (KD), curling arms (CA), swinging (SW) or serving in tennis (ST), swinging a baseball bat (SB), shooting a basketball (SH), drawing a triangle (DT), crossing arms (CA), and drawing an X with the right hand (DX). Sample images depicting these activities have been presented in Figure 11.

2) THE HWU-USP DATASET

Ranieri et al. [46] implemented the HWU-USP dataset at the University of Sao Paulo. Sixteen participants engaged in nine activities within a smart home setting. RGBD data was gathered using the TIAGO robot, while inertial data was acquired by positioning sensors on the wrist and waist. The activities encompass dishwashing (DW), setting table (ST), reading newspaper (RN), using phone (UP), using laptop (UL), cleaning kitchen (CK), making tea (MT), making sandwich (MS), making cereals (MC). The visual representations of sample images have been shown in Figure 13.

3) THE BERKELEY-MHAD DATASET

The Berkeley Multimodal Human Action Database (Berkeley-MHAD) [50] consists of eleven activities executed by a group of 12 participants, consisting of seven males and five females aged between 23 and 30 years. Each activity is repeated five times, resulting in a total of 660 sequences. Experimental actions were captured by camera, accelerometer, and Kinect sensors as depicted in Figure 14. The included activities encompass jumping jack (JJ), bending (BN), sit-down/stand-up (SS), punching (PN), throwing (TH), waving two hands (WT), clapping (CL), standing (ST), jumping (JP), waving one hand (WO), and sitting (SI). These actions were executed in three distinct positions: the lower extreme position, both lower and upper extreme positions, and solely in the upper extreme position, as depicted in Figure 13.

4) THE NTU-RGB+D60 DATASET

Shahroudy et al. [106] collected four major data modalities include depth maps, 3D joint information, RGB frames, and IR sequences. However, in this paper only RGB frames data has been used to evaluate the performance of our model. The dataset consists of 60 action classes including 40 daily actions, 9 health-related actions, and 11 mutual actions. As this paper is related to HAR on daily based actions. So, only 40 daily actions have been used in this paper. The actions were performed by 40 subjects. Their ages lie between 10 and 35. The 40 actions include drink water (DW), eat meal (EM), brush teeth (BT), brush hair (BH), drop (DR), pick up (PU), throw (TH), sit down (SD), stand up (SU), clapping (CL), reading (RG), writing (WG), tear up paper (TP), put on jacket (PJ), take off jacket (TJ), put on a shoe (PS), take off a shoe (TS), put on glasses (PG), take off glasses (TG), put on a hat/cap (PC), take off a hat/cap (TC), cheer up (CU), hand waving (HW), kicking something (KS), reach into pocket (RP), hopping (HG), jump up (JU), phone call (PC), play with phone/tablet (PT), type on a keyboard (TK), point to something (PS), taking a selfie (TS), check time (from watch) (CT), rub two hands (RH), nod head/bow (NB), shake head

FIGURE 11. The 27 sample activities performed in HWU-USP dataset.

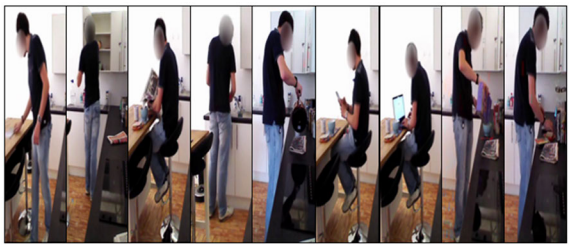


FIGURE 12. The 9 sample activities performed in HWU-USP dataset.

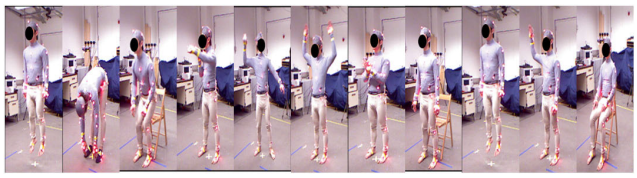


FIGURE 13. The 9 sample activities performed in HWU-USP dataset.

(SH), wipe face (WF), salute (SE), put palms together (PT), cross hands in front (CF).

5) THE NTU-RGB+D120 DATASET

Shahroudy et al. [107] collected four major data modalities include depth maps, 3D joint information, RGB frames, and IR sequences. However, in this paper only RGB frames data has been used to evaluate the performance of our model. The dataset consists of 60 action classes including 40 daily actions, 9 health-related actions, and 11 mutual actions. This paper is related to HAR on daily based actions. So, only 40 daily actions have been used in this paper. The actions were performed by 40 subjects. Their ages lie between 10 and 35. The 40 actions include drink water (DW), eat meal (EM), brush teeth (BT), brush hair (BH), drop (DR), pick up (PU), throw (TH), sit down (SD), stand up (SU), clapping (CL), reading (RG), writing (WG), tear up paper (TP), put on jacket (PJ), take off jacket (TJ), put on a shoe (PS), take off a shoe (TS), put on glasses (PG), take off glasses (TG), put on a hat/cap (PC), take off a hat/cap (TC), cheer up (CU), hand waving (HW), kicking something (KS), reach into pocket (RP), hopping (HG), jump up (JU), phone call (PC), play with phone/tablet (PT), type on a keyboard (TK), point to something (PS), taking a selfie (TS), check time (from watch) (CT), rub two hands (RH), nod head/bow (NB), shake head (SH), wipe face (WF), salute (SE), put palms together (PT), cross hands in front (CF).

B. EXPERIMENTAL SETUP

The proposed methodology was implemented through a Python application on an Intel i7-8250 CPU, 64-bit operating system, 1.9GHz processor, and 32GB RAM. For benchmarking, we selected the Berkeley-MHAD, UTD-MHAD, HWU-USP, NTU-RGB+D60, and NTU-RGB+D120 datasets, as they are the only available multimodal datasets with complete silhouette RGB, depth, and inertial sensor data for human action recognition models. The system being assessed underwent testing using the LOS cross-validation method, employing distinct training and testing datasets. To distinguish between various postures and movements, human activity classification was performed on UTD-MHAD, Berkeley-MHAD, HWU-USP, NTU-RGB+D60, and NTU-RGB+D120 datasets. System evaluation was conducted utilizing Accuracy, Sensitivity, and F1 score as (5), (6) and (7).

$$Precision = \frac{TD}{TD + FD} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + NM} \tag{6}$$

$$F1 = \frac{2\,(Accuracy \times Sensitivity)}{(Accuracy + Sensitvity)} \tag{7}$$

where TD depicts True Detection, FD represents False Detection, TP depicts True Positive and NM epresents Negative Match.

Tables 2, 3, 4 and 5 show the confusion matrices for the HWU-USP, Berkeley-MHAD, NTU-RGB+D60, and NTU-RGB+D120 benchmark datasets, providing a detailed overview of the classification performance. The results indicate how well the proposed model distinguishes between different classes within these datasets. According to Tables 6-10 the system achieves an impressive average F-Score accuracy, reflecting a strong balance between sensitivity and precision in its predictions. This high F-Score underscores the model's effectiveness in handling various classification challenges. Table 11 and Table 12 offers a comparative analysis of our proposed approach against existing framework techniques. This comparison highlights

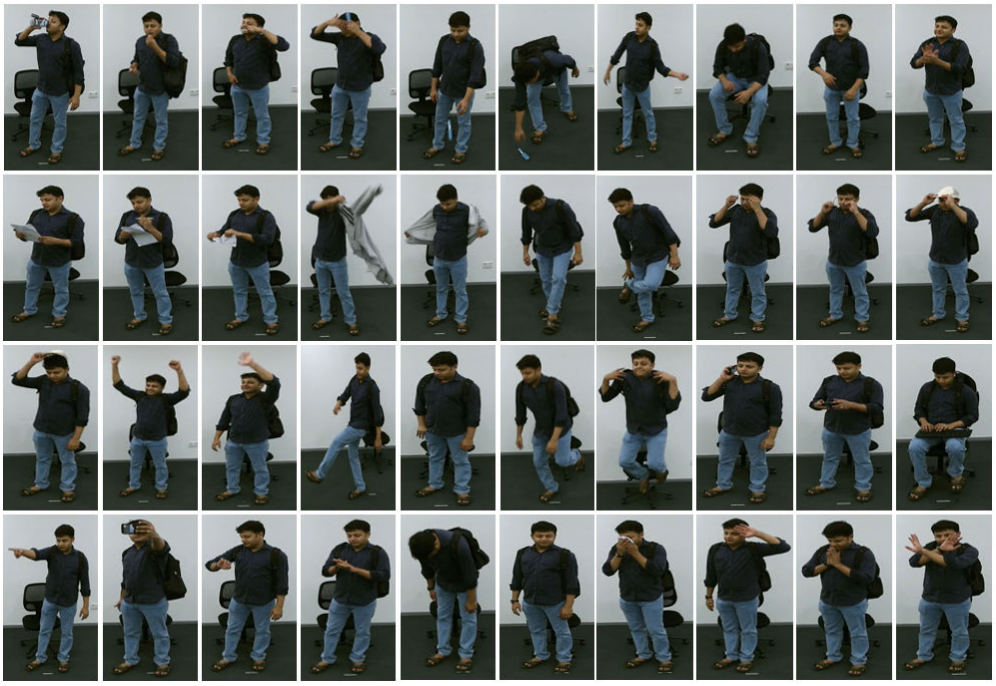


FIGURE 14. The 40 sample activities performed in the NTU-RGB+D60 dataset.

FIGURE 15. The 82 sample activities performed in the NTU-RGB+D120 dataset.

that our strategy not only meets but exceeds the performance of current state-of-the-art methods in relevant domains, demonstrating its superiority and potential for advancing the field.

V. ABLATION STUDY

The ablation study in Table 13 evaluates our model's performance by systematically removing each of its key components. Each row represents a model variation with

TABLE 2. Assessing the classification performance of the HWU-USP dataset using a confusion matrix.

Activity	DW	ST	RN	UP	UL	CK	MT	MS	MC
DW	97.92	0.1	0.08	0.1	0.1	0.5	0.5	0.4	0.3
ST	0.14	96.31	0.7	0.5	1.88	0.41	0.03	0.02	0.01
RN	0.2	0.2	97.82	0.54	0.5	0.1	0.14	0.2	0.3
UP	0.1	0.1	0.5	97.93	0.57	0.3	0.2	0.1	0.2
UL	0.2	0.2	0.5	0.52	97.70	0.26	0.12	0.3	0.2
CK	0.5	0.5	0.03	0.04	0.01	97.91	0.5	0.5	0.01
MT	0.0	0.1	0.0	0.08	0.0	0.0	98.82	0.5	0.5
MS	0.2	0.3	0.2	0.07	0.0	0.5	0.0	98.73	0.0
MC	0.1	0.1	0.1	0.2	0.1	0.3	0.2	0.1	98.80
Mean Accuracy = 97.99%									

TABLE 3. Assessing the classification performance of the Berkeley-MHAD dataset using a confusion matrix.

Activity	JJ	BN	SS	PN	TH	WT	CL	ST	JP	WO	SI
JJ	97.99	0.1	0.2	0.1	0.5	0.0	0.0	0.1	0.91	0.0	0.1
BN	0.02	98.98	0.3	0.0	0.0	0.0	0.0	0.3	0.1	0.0	0.3
SS	0.3	0.2	96.84	0.06	0.3	0.2	0.3	0.7	0.2	0.2	0.7
PN	0.9	0.2	0.1	95.88	0.8	0.3	0.5	0.1	0.7	0.5	0.02
TH	0.2	0.3	0.2	0.4	97.53	0.4	0.17	0.3	0.2	0.2	0.1
WT	0.0	0.0	0.0	0.0	0.1	98.93	0.07	0.0	0.0	0.9	0.0
CL	0.0	0.2	0.0	0.2	0.2	0.4	98.77	0.0	0.03	0.2	0.0
ST	0.0	0.4	0.8	0.0	0.05	0.0	0.0	97.85	0.0	0.0	0.9
JP	1.03	0.0	0.1	0.3	0.1	0.1	0.0	0.0	98.27	0.1	0.0
WO	0.0	0.0	0.0	0.2	0.3	1.01	0.4	0.0	0.1	97.99	0.0
SI	0.1	0.2	0.7	0.0	0.0	0.1	0.3	0.52	0.0	0.1	97.98
Mean Accuracy = 97.91%											

TABLE 4. Assessing the classification performance of the NTU-RGB+D60 dataset using accuracy of individual activity.

Activities	Accuracy (%)	Activities	Accuracy (%)	Activities	Accuracy (%)
DW	95.99%	TJ	98.06%	PT	98.06%
EM	96.53%	PS	99.77%	TK	99.77%
BT	97.84%	TS	96.64%	PS	96.64%
BH	97.32%	PG	97.34%	TS	97.34%
DR	95.41%	TG	94.87%	CT	94.87%
PU	98.07%	PC	95.12%	RH	95.12%
TH	96.13%	TC	98.32%	NB	98.32%
SD	95.99%	CU	97.09%	SH	97.09%
SU	97.08%	HW	96.75%	WF	96.75%
CL	96.45%	KS	97.84%	SE	97.84%
RG	97.02%	RP	97.36%	PT	97.36%
WG	94.99%	HG	96.55%	CF	96.55%
TP	97.58%	JU	98.15%		
PJ	94.87%	PC	97.13%		
Mean Accuracy = 96.61%					

a specific component removed, and the corresponding accuracy is tested across the following datasets: NTU-RGB+D60, NTU-RGB+D120, Berkeley-MHAD, HWU-USP, and UTD-MHAD. This Table highlights the importance of each component in achieving high accuracy. The study demonstrates the effectiveness and robustness of the proposed model elements, as the removal of certain component—such as CNN-GRU, Lagregression (LR), Gammatone Cepstral Coefficients (GCC), Dynamic Likelihood Random Field (DLRF), Angle along the Sagittal Plane (ASP), and the GA that significantly degrades performance, underscoring their critical contributions. Notably, the most substantial decrease in accuracy occurs when GCC and CNN-GRU is excluded, emphasizing its pivotal role in the detection pipeline. Additionally, a noticeable decline in performance is observed when the GA is removed, highlighting the importance of

TABLE 5. Assessing the classification performance of the NTU-RGB+D120 dataset using accuracy of individual activity.

Activities	Accuracy (%)	Activities	Accuracy (%)	Activities	Accuracy (%)
DW	94.19%	PT	98.18%	SF	98.23%
EM	95.23%	TK	96.23%	OB	97.41%
BT	98.17%	PS	97.14%	SS	97.64%
BH	93.05%	TS	97.44%	SD	99.79%
DR	95.11%	CT	98.12%	TC	90.89%
PU	94.07%	RH	93.27%	FP	97.16%
TH	92.99%	NB	98.37%	BP	95.92%
SD	97.12%	SH	95.01%	PM	99.79%
SU	99.89%	WF	93.07%	AF	98.17%
CL	92.05%	SE	99.82%	AH	94.91%
RG	96.88%	PT	93.02%	PB	95.32%
WG	94.03%	CF	99.11%	TB	98.71%
TP	99.03%	PH	98.34%	PO	93.55%
PJ	94.12%	TH	95.72%	TB	95.11%
TJ	93.75%	SB	93.22%	OB	94.99%
PS	90.18%	BB	98.73%	MO	97.12%
TS	93.02%	TS	90.98%	SF	91.98%
PG	97.64%	JB	96.35%	TC	92.21%
TG	94.24%	HH	97.18%	CE	98.81%
PC	95.13%	FH	92.11%	CA	99.85%
TC	96.34%	TU	98.03%	AC	93.46%
CU	94.88%	TD	94.33%	AS	99.18%
HW	98.02%	MS	93.57%	RS	93.75%
KS	99.58%	MV	92.09%	BK	98.13%
RP	93.64%	SB	98.31%	CH	97.72%
HG	97.11%	CM	97.27%	SK	98.37%
JU	99.81%	CN	93.07%		
PC	96.02%	CP	96.23%		
Mean Accuracy = 95.94%					

TABLE 6. Evaluating the classification performance of HWU-USP dataset using accuracy, sensitivity, and F-score.

Class	Accuracy	Sensitivity	F-Score	Class	Accuracy	Sensitivity	F-score
DW	0.979	0.978	0.978	CK	0.979	0.978	0.978
ST	0.963	0.964	0.963	MT	0.988	0.989	0.988
RN	0.978	0.979	0.978	MS	0.987	0.986	0.986
UP	0.979	0.978	0.978	MC	0.988	0.987	0.987
UL	0.977	0.978	0.977				
Mean Precision = 0.979			**Mean Recall = 0.979**			**Mean F-Score = 0.979**	

employing a sophisticated feature fusion method. These findings confirm that the integrated components are essential, and their combined effect is what enables the model to achieve high accuracy.

VI. DISCUSSION

The paper introduces a novel multimodal framework for human action recognition designed to assist individuals with daily activities. The model's performance is highly dependent on the quality and quantity of input data for training. Using existing benchmark datasets, the model achieved an average accuracy of 96% across five benchmark datasets. The suboptimal performance in certain activities is attributed to data imbalances and insufficient amounts in the existing datasets. However, the model is capable of adapting to unseen data over time, ensuring its long-term adaptability. Additionally, the proposed model is compared with established

TABLE 7. Evaluating the classification performance of Berkeley-MHAD dataset using accuracy, sensitivity, and F-score.

Class	Accuracy	Sensitivity	F-Score	Class	Accuracy	Sensitivity	F-score
JJ	0.979	0.978	0.978	CL	0.987	0.988	0.987
BN	0.989	0.988	0.988	ST	0.978	0.979	0.978
SS	0.968	0.969	0.968	JP	0.982	0.981	0.981
PN	0.958	0.957	0.957	WO	0.979	0.978	0.978
TH	0.975	0.976	0.975	SI	0.979	0.978	0.978
WT	0.989	0.988	0.988				
		Mean Precision = 0.979		**Mean Recall =0.979**		**Mean F-Score =0.979**	

TABLE 8. The determination of classification performance of the UTD-MHAD dataset via precision, recall, and F-score.

Class	Accuracy	Sensitivity	F-Score	Class	Accuracy	Sensitivity	F-score
SL	0.957	0.958	0.957	SQ	0.981	0.980	0.980
SR	0.978	0.979	0.978	WK	0.946	0.947	0.946
DC	0.990	0.991	0.990	TH	0.997	0.996	0.996
DA	0.983	0.984	0.983	PR	0.996	0.997	0.996
SI	0.986	0.985	0.985	FL	0.984	0.985	0.984
ST	0.999	0.998	0.998	KD	0.975	0.976	0.975
WR	0.992	0.993	0.992	CA	0.983	0.984	0.983
CL	0.977	0.978	0.977	SW	0.956	0.955	0.955
BW	0.990	0.991	0.990	ST	0.993	0.992	0.992
BX	0.988	0.989	0.988	SB	0.981	0.982	0.981
PU	0.979	0.978	0.978	SH	0.947	0.948	0.947
CT	0.970	0.971	0.970	DT	0.996	0.997	0.996
TR	0.956	0.957	0.956	CA	0.987	0.988	0.987
JG	0.964	0.965	0.964	DX	0.995	0.996	0.995
	Mean Accuracy = 0.979		**Mean Recall = 0.979**		**Mean F-Score = 0.979**		

TABLE 9. Evaluating the classification performance of NTU-RGB+D60 dataset using precision, sensitivity, and F-score.

Class	Precision	Sensitivity	F-Score	Class	Precision	Sensitivity	F-score
DW	92.10%	92.00%	92.00%	TC	95.80%	93.80%	94.70%
EM	94.50%	93.90%	93.90%	CU	90.30%	90.80%	90.50%
BT	90.10%	89.90%	89.90%	HW	91.70%	91.20%	91.40%
BH	93.20%	93.60%	93.60%	KS	92.20%	92.50%	92.30%
DR	90.50%	90.30%	90.30%	RP	91.90%	91.30%	91.60%
PU	92.40%	92.00%	92.00%	HG	93.60%	93.20%	93.40%
TH	93.10%	93.30%	93.30%	JU	90.80%	91.50%	91.10%
SD	91.40%	91.20%	91.20%	PC	91.20%	90.80%	91.00%
SU	92.70%	92.50%	92.50%	PT	94.40%	93.50%	93.90%
CL	92.30%	92.20%	92.20%	TK	91.30%	92.50%	91.90%
RG	93.90%	93.60%	93.60%	PS	92.20%	91.40%	91.80%
WG	93.20%	93.00%	93.00%	TS	91.60%	90.90%	91.20%
TP	90.90%	94.70%	94.70%	CT	91.20%	90.90%	91.00%
PJ	91.70%	91.20%	91.20%	RH	95.40%	95.00%	95.20%
TJ	93.80%	93.30%	93.30%	NB	92.30%	91.80%	92.00%
PS	92.50%	92.20%	92.20%	SH	90.20%	91.50%	90.80%
TS	91.20%	90.90%	90.90%	WF	91.10%	90.90%	91.00%
PG	90.10%	90.30%	90.30%	SE	90.00%	91.20%	90.60%
TG	91.80%	91.30%	91.30%	PT	90.10%	91.60%	90.80%
PC	93.50%	93.10%	93.10%	CF	93.80%	92.20%	92.90%
Mean Accuracy = 92.15%			**Mean Sensitivity = 90.70%**			**Mean F1-Score = 92.00%**	

TABLE 10. Evaluating the classification performance of NTU-RGB+D120 dataset using accuracy, sensitivity, and F-score.

Class	Accuracy	Sensitivity	F-Score	Class	Accuracy	Sensitivity	F-score
DW	94.10%	93.60%	93.80%	TH	93.90%	92.70%	93.30%
EM	93.80%	92.70%	93.20%	SB	91.00%	90.70%	90.80%
BT	94.70%	93.60%	94.10%	BB	91.70%	90.40%	91.00%
BH	94.20%	94.80%	94.50%	TS	97.50%	96.40%	96.90%
DR	95.80%	96.90%	96.30%	JB	93.50%	92.70%	93.10%
PU	90.90%	90.30%	90.60%	HH	94.30%	93.20%	93.70%
TH	92.10%	92.50%	92.30%	FH	91.20%	90.10%	90.60%
SD	99.70%	98.60%	99.10%	TU	99.80%	97.60%	98.60%
SU	96.80%	96.60%	96.70%	TD	93.70%	92.40%	93.00%
CL	94.30%	93.20%	93.70%	MS	91.40%	90.30%	90.80%
RG	93.50%	94.80%	94.10%	MV	93.40%	92.30%	92.80%
WG	92.90%	91.90%	92.40%	SB	90.10%	99.10%	94.30%
TP	95.70%	94.60%	95.10%	CM	91.10%	98.20%	94.50%
PJ	97.10%	96.80%	96.90%	CN	93.10%	92.90%	93.00%
TJ	90.10%	91.10%	90.60%	CP	90.60%	91.50%	91.00%
PS	93.90%	93.50%	93.70%	SF	90.00%	90.90%	90.40%
TS	91.10%	90.10%	90.60%	OB	93.80%	91.70%	92.70%
PG	98.10%	91.70%	94.70%	SS	94.10%	93.80%	93.90%
TG	97.00%	96.60%	96.80%	SD	91.10%	90.70%	90.90%
PC	90.40%	93.40%	91.80%	TC	93.40%	92.30%	92.80%
TC	91.90%	90.80%	91.30%	FP	91.20%	90.40%	90.80%
CU	90.70%	91.30%	91.00%	BP	99.10%	98.00%	98.50%
HW	91.10%	90.80%	90.90%	PM	93.10%	91.20%	92.10%
KS	93.10%	92.80%	92.90%	AF	91.20%	90.80%	91.00%
RP	90.50%	91.70%	91.10%	AH	97.10%	96.80%	96.90%
HG	93.30%	93.50%	93.40%	PB	90.40%	91.10%	90.70%
JU	92.90%	91.70%	92.30%	TB	91.30%	90.70%	91.00%
PC	91.10%	91.90%	91.50%	PO	91.00%	90.10%	90.50%
PT	90.60%	90.10%	90.30%	TB	91.80%	91.30%	91.50%
TK	91.30%	91.10%	91.20%	OB	96.80%	94.50%	95.60%
PS	96.20%	94.50%	95.30%	MO	91.40%	93.80%	92.50%
TS	90.50%	91.90%	91.10%	SF	95.10%	94.80%	94.90%
CT	91.70%	91.40%	91.50%	TC	98.30%	97.10%	97.70%
RH	92.20%	90.30%	91.20%	CE	92.10%	91.90%	92.00%
NB	91.30%	90.90%	91.10%	CA	93.10%	94.80%	93.90%
SH	98.80%	97.40%	98.10%	AC	92.70%	91.80%	92.20%
WF	92.10%	91.60%	91.80%	AS	93.60%	90.40%	91.90%
SE	94.30%	93.20%	93.70%	RS	93.10%	92.80%	92.90%
PT	99.90%	95.60%	97.70%	BK	94.70%	93.80%	94.20%
CF	91.40%	90.50%	90.90%	CH	90.70%	91.90%	91.30%
PH	93.80%	91.70%	92.70%	SK	90.10%	91.10%	90.60%
Mean Accuracy = 93.30%			**Mean Sensitivity = 92.90%**			**Mean F1-Score = 93.00%**	

TABLE 11. Assessing the proposed approach compared to other state-of-the-art methods.

Authors	Methodology	Accuracy
Javeed et al. [43]	Time and Probability Domain Features	82.50%
Cao et al. [108]	Modified Dynamic Time Warping (MDTW)	89.50%
Cohen et al. [109]	2D CNN & 3D CNN	86.20%
Dhiman et al. [110]	Deep View-Invariant HAR Framework	87.30%
Proposed	**Novel CNNGRU**	**97.99%**

machine learning and deep learning techniques, as detailed in Table 11-12.

The inertial data has been divided into 5-second windows, while both RGB and Depth images were segmented into

TABLE 12. Assessing the proposed approach compared to other state-of-the-art methods.

Authors	Methodology	Datasets	Accuracy
Yang [41]	Depth Sequential Information Entropy Maps	UTD-MHAD	88.37%
Golestani [42]	Random Forest	Berkeley-MHAD	91.50%
Javeed [43]	Recurrent Neural Network	HWU-USP	87.67%
Kang [44]	CNN+LSTM-RNN	Berkeley-MHAD	94.80%
Ni [45]	Progressive Skeleton-to-sensor Knowledge Distillation (PSKD)	UTD-MHAD	95.19%
Ranieri [46]	Deep Learning Framework	HWU-USP	93.75%
Proposed	**Novel CNNGRU**	**UTD-MHAD, Berkeley-MHAD, HWU-USP**	**97.99%**

TABLE 13. Ablation study on NTU-RGB+D60, NTU-RGB+D120, Berkeley-MHAD, HWU-USP, and UTD-MHAD.

Experiments	LR	GCC	DLRF	ASP	GA	CNN-GRU	NTU-RGB+D60	NTU-RGB+D120	Berkeley-MHAD	HWU-USP	UTD-MHAD
Full architecture	O	O	O	O	O	O	96.61%	95.94%	97.91%	97.99%	97.90%
w/o LR	×	O	O	O	O	O	95.09%	94.11%	96.88%	96.15%	96.33%
w/o GCC	O	×	O	O	O	O	90.02%	90.08%	91.12%	91.38%	91.09%
w/o DLRF	O	O	×	O	O	O	94.11%	94.09%	95.17%	95.28%	95.14%
w/o ASP	O	O	O	×	O	O	95.18%	94.45%	95.99%	95.18%	94.68%
w/o GA	O	O	O	O	×	O	90.78%	90.14%	91.08%	92.34%	91.99%
w/o CNN-GRU	O	O	O	O	O	×	87.49%	87.66%	89.09%	89.57%	88.39%
w/o RGB-D features	O	O	×	×	O	O	91.58%	91.19%	92.48%	92.66%	92.27%
w/o Inertial Features	×	×	O	O	O	O	92.33%	92.45%	93.48%	93.56%	93.09%

9-second frames. Using a Kaiser window low-pass filter, 500 samples were obtained. Simultaneously, the silhouette extraction process for vision sensor data produced 240 binary images. During the feature extraction phase, a total of 6,290 sample data points were extracted. Finally, the novel CNNGRU classifier provided classification results for 1,735 sample data points.

The computational complexity of the proposed CNNGRU model per weight and time step is O(1), and the computational complexity per time step is $zO(N)$. Additionally, the average learning time is primarily influenced by the number of memory cells m_c and the number of outputs m_0, calculated as $m_c(4m_c+m_0)$. However, the model becomes computationally expensive when the number of output units increases significantly to store temporal contextual information. Despite this, the model shows no effectiveness in recognizing complex tasks on benchmark datasets compared to traditional models.

VII. CONCLUSION AND FUTURE DIRECTION

The framework utilizes multimodal sensor data for human action recognition and integrates modified features for inertial, depth, and RGB sensor data. These improved features, including DLRF, AR, and GCC, are employed for recognizing complex human action patterns, providing both contextual information and behavior classification. The extracted multimodal features are subsequently fused using a GA. Furthermore, a novel CNNGRU classifier is proposed for the classification of various human patterns. Our proposed system achieves an average accuracy of 96% over benchmark datasets, showcasing its potential for real-world applications. This study holds promise in reducing the time, costs, and errors associated with diagnosing complex and diverse patterns of human daily life activities.

In the future, we plan to develop our own benchmark multimodal datasets, including RGB, depth, and inertial sensor data. Additionally, we aim to introduce new features to enhance the human action recognition framework and provide real-time support for daily activities. These contributions are expected to significantly improve the model's applicability and positively impact daily human activities.

REFERENCES

[1] M. B. Shaikh, S. M. S. Islam, D. Chai, and N. Akhtar, "From CNNs to transformers in multimodal human action recognition: A survey," 2024, *arXiv:2405.15813*.

[2] D. Khan, M. Alonazi, M. Abdelhaq, N. A. Mudawi, A. Algarni, A. Jalal, and H. Liu, "Robust human locomotion and localization activity recognition over multisensory," *Frontiers Physiol.*, vol. 15, Feb. 2024, Art. no. 1344887, doi: 10.3389/fphys.2024.1344887.

[3] K. Wang, A. Boonpratatong, W. Chen, L. Ren, G. Wei, Z. Qian, X. Lu, and D. Zhao, "The fundamental property of human leg during walking: Linearity and nonlinearity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 4871–4881, 2023, doi: 10.1109/TNSRE.2023.3339801.

[4] Y. Zhou, J. Xie, X. Zhang, W. Wu, and S. Kwong, "Energy-efficient and interpretable multisensor human activity recognition via deep fused Lasso net," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 5, pp. 3576–3588, Oct. 2024, doi: 10.1109/TETCI.2024.3430008.

[5] J. Ding, X. Chen, P. Lu, Z. Yang, X. Li, and Y. Du, "DialogueINAB: An interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition," *J. Supercomput.*, vol. 79, no. 18, pp. 20481–20514, Dec. 2023, doi: 10.1007/s11227-023-05439-1.

[6] H.-B. Huang, Y.-L. Zheng, and Z.-Y. Hu, "Video abnormal action recognition based on multimodal heterogeneous transfer learning," *Adv. Multimedia*, vol. 2024, Jan. 2024, Art. no. 4187991.

[7] C. Gupta, N. S. Gill, P. Gulia, S. Yadav, G. Pau, M. Alibakhshikenari, and X. Kong, "A real-time 3-dimensional object detection based human action recognition model," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 14–26, 2024.

[8] X. Wang, R. Zhang, Y. Miao, M. An, S. Wang, and Y. Zhang, "PI^2-based adaptive impedance control for gait adaption of lower limb exoskeleton," *IEEE/ASME Trans. Mechatron.*, early access, Mar. 18, 2024, doi: 10.1109/TMECH.2024.3370954.

[9] R. Zhang, L. Li, Q. Zhang, J. Zhang, L. Xu, B. Zhang, and B. Wang, "Differential feature awareness network within antagonistic learning for infrared-visible object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6735–6748, Aug. 2023, doi: 10.1109/TCSVT.2023.3289142.

[10] C. Hu, B. Dong, H. Shao, J. Zhang, and Y. Wang, "Toward purifying defect feature for multilabel sewer defect classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023, doi: 10.1109/TIM.2023.3250306.

[11] H. Niu, D. Nguyen, K. Yonekawa, M. Kurokawa, S. Wada, and K. Yoshihara, "Multi-source transfer learning for human activity recognition in smart homes," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Bologna, Italy, Sep. 2020, pp. 274–277, doi: 10.1109/SMARTCOMP50058.2020.00063.

[12] S. Wu, S. Zhang, D. Xu, and D. Huang, "A weighting localization algorithm with LOS and one-bound NLOS identification in multipath environments," *J. Inf. Sci. Eng.*, vol. 35, no. 6, pp. 1209–1222, 2019, doi: 10.6688/JISE.201911_35(6).0003.

[13] S. He, W. Chen, K. Wang, H. Luo, F. Wang, W. Jiang, and H. Ding, "Region generation and assessment network for occluded person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 120–132, 2024, doi: 10.1109/TIFS.2023.3318956.

[14] S. Nanda and S. K. Dutta, "Assessing human activity recognition performances of different machine learning algorithms using sensor data," in *Proc. IEEE Silchar Subsection Conf. (SILCON)*, Silchar, India, Nov. 2023, pp. 1–6, doi: 10.1109/silcon59133.2023.10404163.

[15] J. Xing, H. Yuan, R. Hamzaoui, H. Liu, and J. Hou, "GQE-Net: A graph-based quality enhancement network for point cloud color attribute," *IEEE Trans. Image Process.*, vol. 32, pp. 6303–6317, 2023, doi: 10.1109/TIP.2023.3330086.

[16] X. Gu, X. Chen, P. Lu, X. Lan, X. Li, and Y. Du, "SiMaLSTM-SNP: Novel semantic relatedness learning model preserving both Siamese networks and membrane computing," *J. Supercomput.*, vol. 80, no. 3, pp. 3382–3411, Feb. 2024, doi: 10.1007/s11227-023-05592-7.

[17] Z. Zhou, Y. Wang, R. Liu, C. Wei, H. Du, and C. Yin, "Short-term lateral behavior reasoning for target vehicles considering driver preview characteristic," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11801–11810, Aug. 2022, doi: 10.1109/TITS.2021.3107310.

[18] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan, and M. Zaharadeen, "Automated daily human activity recognition for video surveillance using neural network," in *Proc. IEEE 4th Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, Putrajaya, Malaysia, Nov. 2017, pp. 1–5, doi: 10.1109/ICSIMA.2017.8312024.

[19] M. Javeed and A. Jalal, "Deep activity recognition based on patterns discovery for healthcare monitoring," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Lahore, Pakistan, Feb. 2023, pp. 1–6, doi: 10.1109/ICACS55311.2023.10089764.

[20] J. J. Peng, X. G. Chen, X. K. Wang, J. Q. Wang, Q. Q. Long, and L. J. Yin, "Picture fuzzy decision-making theories and methodologies: A systematic review," *Int. J. Syst. Sci.*, vol. 54, no. 13, pp. 2663–2675, Oct. 2023, doi: 10.1080/00207721.2023.2241961.

[21] C. Zhu, "Intelligent robot path planning and navigation based on reinforcement learning and adaptive control," *J. Logistics, Inform. Service Sci.*, vol. 10, no. 3, pp. 235–248, 2023, doi: 10.33168/JLISS.2023.0318.

[22] M. B. Abidine, B. Fergani, and I. Menhour, "Activity recognition from smartphones using hybrid classifier PCA-SVM-HMM," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Fez, Morocco, Oct. 2019, pp. 1–5, doi: 10.1109/WINCOM47513.2019.8942492.

[23] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019, doi: 10.1109/ACCESS.2019.2913393.

[24] H. Nagaraja and Abhishek, "Heart rate and SPO_2 monitoring app while exercising squats for smartphone," in *Proc. 1st Int. Conf. Informat. (ICI)*, Noida, India, Apr. 2022, pp. 244–246, doi: 10.1109/ici53355.2022.9786876.

[25] W. Zheng, S. Lu, Y. Yang, Z. Yin, and L. Yin, "Lightweight transformer image feature extraction network," *PeerJ Comput. Sci.*, vol. 10, p. e1755, Jan. 2024, doi: 10.7717/peerj-cs.1755.

[26] Y. Chen, N. Li, D. Zhu, C. C. Zhou, Z. Hu, Y. Bai, and J. Yan, "BEVSOC: Self-supervised contrastive learning for calibration-free BEV 3-D object detection," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 22167–22182, Jun. 2024, doi: 10.1109/JIOT.2024.3379471.

[27] S. Lu, J. Yang, B. Yang, X. Li, Z. Yin, L. Yin, and W. Zheng, "Surgical instrument posture estimation and tracking based on LSTM," *ICT Exp.*, vol. 10, no. 3, pp. 465–471, Jun. 2024, doi: 10.1016/j.icte.2024.01.002.

[28] P. Deepan, R. Santhosh Kumar, B. Rajalingam, P. S. Kumar Patra, and S. Ponnuthurai, "An intelligent robust one dimensional HAR-CNN model for human activity recognition using wearable sensor data," in *Proc. 4th Int. Conf. Adv. Comput., Commun. Control Netw. (ICAC3N)*, Greater Noida, India, Dec. 2022, pp. 1132–1138, doi: 10.1109/ICAC3N56670.2022.10073991.

[29] A. Jalal, S. Kamal, and D. Kim, "Depth map-based human activity tracking and recognition using body joints features and self-organized map," in *Proc. 5th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Hefei, China, Jul. 2014, pp. 1–6, doi: 10.1109/ICCCNT.2014.6963013.

[30] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023, doi: 10.1109/TPAMI.2022.3183112.

[31] M. H. Siddiqi, N. Almashfi, A. Ali, M. Alruwaili, Y. Alhwaiti, S. Alanazi, and M. M. Kamruzzaman, "A unified approach for patient activity recognition in healthcare using depth camera," *IEEE Access*, vol. 9, pp. 92300–92317, 2021, doi: 10.1109/ACCESS.2021.3092403.

[32] F. Qi, X. Tan, Z. Zhang, M. Chen, Y. Xie, and L. Ma, "Glass makes blurs: Learning the visual blurriness for glass surface detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 4, pp. 6631–6641, Apr. 2024, doi: 10.1109/TII.2024.3352232.

[33] J. Sun, L. Zhou, B. Geng, Y. Zhang, and Y. Li, "Leg state estimation for quadruped robot by using probabilistic model with proprioceptive feedback," *IEEE/ASME Trans. Mechtron.*, early access, Jul. 16, 2024, doi: 10.1109/TMECH.2024.3421251.

[34] S. Pan, G. J. W. Xu, K. Guo, S. H. Park, and H. Ding, "Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach," *IEEE Trans. Games*, early access, Dec. 29, 2023, doi: 10.1109/TG.2023.3348230.

[35] Z. Chen, Q. Liang, Z. Wei, X. Chen, Q. Shi, Z. Yu, and T. Sun, "An overview of in vitro biological neural networks for robot intelligence," *Cyborg Bionic Syst.*, vol. 4, p. 1, Jan. 2023, doi: 10.34133/cbsystems.0001.

[36] C. Zhu, "An adaptive agent decision model based on deep reinforcement learning and autonomous learning," *J. Logistics, Informat. Service Sci.*, vol. 10, no. 3, pp. 107–118, 2023, doi: 10.33168/JLISS.2023.0309.

[37] X. Liu, Y. Wang, Z. Zhou, K. Nam, C. Wei, and C. Yin, "Trajectory prediction of preceding target vehicles based on lane crossing and final points generation model considering driving styles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8720–8730, Sep. 2021, doi: 10.1109/TVT.2021.3098429.

[38] H. He, X. Li, P. Chen, J. Chen, M. Liu, and L. Wu, "Efficiently localizing system anomalies for cloud infrastructures: A novel dynamic graph transformer based parallel framework," *J. Cloud Comput.*, vol. 13, no. 1, Jun. 2024, Art. no. 115, doi: 10.1186/s13677-024-00677-x.

[39] X. Shen, H. Jiang, D. Liu, K. Yang, F. Deng, J. C. S. Lui, J. Liu, S. Dustdar, and J. Luo, "PupilRec: Leveraging pupil morphology for recommending on smartphones," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 15538–15553, Sep. 2022, doi: 10.1109/JIOT.2022.3181607.

[40] Q. Wang, J. Hu, Y. Wu, and Y. Zhao, "Output synchronization of wide-area heterogeneous multi-agent systems over intermittent clustered networks," *Inf. Sci.*, vol. 619, pp. 263–275, Jan. 2023, doi: 10.1016/j.ins.2022.11.035.

[41] T. Yang, Z. Hou, J. Liang, Y. Gu, and X. Chao, "Depth sequential information entropy maps and multi-label subspace learning for human action recognition," *IEEE Access*, vol. 8, pp. 135118–135130, 2020.

[42] N. Golestani and M. Moghaddam, "A comparison of machine learning classifiers for human activity recognition using magnetic induction-based motion signals," in *Proc. 14th Eur. Conf. Antennas Propag. (EuCAP)*, vol. 59, Mar. 2020, pp. 1–3.

[43] M. Javeed, M. Shorfuzzaman, N. Alsufyani, S. A. Chelloug, A. Jalal, and J. Park, "Physical human locomotion prediction using manifold regularization," *PeerJ Comput. Sci.*, vol. 8, p. e1105, Oct. 2022.

[44] J. Kang, J. Shin, J. Shin, D. Lee, and A. Choi, "Robust human activity recognition by integrating image and accelerometer sensor data using deep fusion network," *Sensors*, vol. 22, no. 1, p. 174, Dec. 2021.

[45] J. Ni, A. H. H. Ngu, and Y. Yan, "Progressive cross-modal knowledge distillation for human action recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5903–5912.

[46] C. M. Ranieri, S. MacLeod, M. Dragone, P. A. Vargas, and R. A. F. Romero, "Activity recognition for ambient assisted living with videos, inertial units and ambient sensors," *Sensors*, vol. 21, no. 3, p. 768, Jan. 2021.

[47] P. William, G. R. Lanke, D. Bordoloi, A. Shrivastava, A. P. Srivastavaa, and S. V. Deshmukh, "Assessment of human activity recognition based on impact of feature extraction prediction accuracy," in *Proc. 4th Int. Conf. Intell. Eng. Manage. (ICIEM)*, May 2023, pp. 1–6.

[48] R. Khan, M. Abbas, R. Anjum, F. Waheed, S. Ahmed, and F. Bangash, "Evaluating machine learning techniques on human activity recognition using accelerometer data," in *Proc. Int. Conf. U.K.-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–6.

[49] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 168–172, doi: 10.1109/ICIP.2015.7350781.

[50] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Clearwater Beach, FL, USA, Jan. 2013, pp. 53–60, doi: 10.1109/WACV.2013.6474999.

[51] I. Alrashdi, M. H. Siddiqi, Y. Alhwaiti, M. Alruwaili, and M. Azad, "Maximum entropy Markov model for human activity recognition using depth camera," *IEEE Access*, vol. 9, pp. 160635–160645, 2021, doi: 10.1109/ACCESS.2021.3132559.

[52] Y. Shi, X. Hou, Z. Na, J. Zhou, N. Yu, S. Liu, L. Xin, G. Gao, and Y. Liu, "Bio-inspired attachment mechanism of dynastes hercules: Vertical climbing for on-orbit assembly legged robots," *J. Bionic Eng.*, vol. 21, no. 1, pp. 137–148, Jan. 2024, doi: 10.1007/s42235-023-00423-0.

[53] Y. Liu, Z. Jia, Z. Jiang, X. Lin, J. Liu, Q. Wu, and W. Susilo, "BFL-SA: Blockchain-based federated learning via enhanced secure aggregation," *J. Syst. Archit.*, vol. 152, Jul. 2024, Art. no. 103163, doi: 10.1016/j.sysarc.2024.103163.

[54] Z. Ahmad and N. Khan, "Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors," *IEEE Sensors J.*, vol. 20, no. 3, pp. 1445–1455, Feb. 2020, doi: 10.1109/JSEN.2019.2947446.

[55] S. Zhang, C. Wang, H. Zhang, and H. Lin, "Collective dynamics of adaptive memristor synapse-cascaded neural networks based on energy flow," *Chaos, Solitons Fractals*, vol. 186, Sep. 2024, Art. no. 115191, doi: 10.1016/j.chaos.2024.115191.

[56] S. Lu, J. Yang, B. Yang, Z. Yin, M. Liu, L. Yin, and W. Zheng, "Analysis and design of surgical instrument localization algorithm," *Comput. Model. Eng. Sci.*, vol. 137, no. 1, pp. 669–685, 2023, doi: 10.32604/cmes.2023.027417.

[57] Y. Tian and W. Chen, "MEMS-based human activity recognition using smartphone," in *Proc. 35th Chin. Control Conf. (CCC)*, Chengdu, China, Jul. 2016, pp. 3984–3989, doi: 10.1109/ChiCC.2016.7553975.

[58] S. Mekruksavanich and A. Jitpattanakul, "Exercise activity recognition with surface electromyography sensor using machine learning approach," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. With ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng. (ECTI DAMT NCON)*, Pattaya, Thailand, Mar. 2020, pp. 75–78, doi: 10.1109/ECTIDAMTNCON48261.2020.9090711.

[59] F. An, J. Wang, and R. Liu, "Road traffic sign recognition algorithm based on cascade attention-modulation fusion mechanism," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 14, 2024, doi: 10.1109/TITS.2024.3439699.

[60] Q. Gao, Z. Deng, Z. Ju, and T. Zhang, "Dual-hand motion capture by using biological inspiration for bionic bimanual robot teleoperation," *Cyborg Bionic Syst.*, vol. 4, Jan. 2023, Art. no. 52, doi: 10.34133/cbsystems.0052.

[61] S. Cristina, V. Despotovic, R. Pérez-Rodríguez, and S. Aleksic, "Audio- and video-based human activity recognition systems in healthcare," *IEEE Access*, vol. 12, pp. 8230–8245, 2024, doi: 10.1109/ACCESS.2024.3353138.

[62] D. Wang, W. Zhang, W. Wu, and X. Guo, "Soft-label for multi-domain fake news detection," *IEEE Access*, vol. 11, pp. 98596–98606, 2023, doi: 10.1109/ACCESS.2023.3313602.

[63] Y. Ding, W. Zhang, X. Zhou, Q. Liao, Q. Luo, and L. M. Ni, "FraudTrip: Taxi fraudulent trip detection from corresponding trajectories," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12505–12517, Aug. 2021, doi: 10.1109/JIOT.2020.3019398.

[64] K. K. Krishnaprabha and C. K. Raju, "Predicting human activity from mobile sensor data using CNN architecture," in *Proc. Adv. Comput. Commun. Technol. High Perform. Appl. (ACCTHPA)*, Jul. 2020, pp. 206–210.

[65] M. V. C. Caya, A. N. Yumang, J. V. Arai, J. D. A. Niñofranco, and K. A. S. Yap, "Human activity recognition based on accelerometer vibrations using artificial neural network," in *Proc. IEEE 11th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ., Manage. (HNICEM)*, Nov. 2019, pp. 1–5.

[66] M. Mahmood, A. Jalal, and K. Kim, "WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors," *Multimedia Tools Appl.*, vol. 79, nos. 11–12, pp. 6919–6950, Mar. 2020, doi: 10.1007/s11042-019-08527-8.

[67] A. Jalal, N. Khalid, and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy Markov model using depth sensors," *Entropy*, vol. 22, no. 8, p. 817, Jul. 2020, doi: 10.3390/e22080817.

[68] M. Pervaiz and A. Jalal, "Artificial neural network for human object interaction system over aerial images," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Lahore, Pakistan, Feb. 2023, pp. 1–6, doi: 10.1109/ICACS55311.2023.10089722.

[69] N. Khalid, M. Gochoo, A. Jalal, and K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, p. 970, Jan. 2021, doi: 10.3390/su13020970.

[70] H. Ansar, A. Jalal, M. Gochoo, and K. Kim, "Hand gesture recognition based on auto-landmark localization and reweighted genetic algorithm for healthcare muscle activities," *Sustainability*, vol. 13, no. 5, p. 2961, Mar. 2021, doi: 10.3390/su13052961.

[71] M. Waheed, S. A. Chelloug, M. Shorfuzzaman, A. Alsufyani, A. Jalal, K. Alnowaiser, and J. Park, "Exploiting human pose and scene information for interaction detection," *Comput., Mater. Continua*, vol. 74, no. 3, pp. 5853–5870, 2023, doi: 10.32604/cmc.2023.033769.

[72] U. Azmat, S. S. Alotaibi, N. A. Mudawi, B. I. Alabduallah, M. Alonazi, A. Jalal, and J. Park, "An elliptical modeling supported system for human action deep recognition over aerial surveillance," *IEEE Access*, vol. 11, pp. 75671–75685, 2023, doi: 10.1109/ACCESS.2023.3266774.

[73] A. Alazeb, U. Azmat, N. A. Mudawi, A. Alshahrani, S. S. Alotaibi, N. A. Almujally, and A. Jalal, "Intelligent localization and deep human activity recognition through IoT devices," *Sensors*, vol. 23, no. 17, p. 7363, Aug. 2023, doi: 10.3390/s23177363.

[74] S. Xia, L. Chu, L. Pei, Z. Zhang, W. Yu, and R. C. Qiu, "Learning disentangled representation for mixed-reality human activity recognition with a single IMU sensor," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[75] S. Mekruksavanich, P. Jantawong, and A. Jitpattanakul, "Deep learning approaches for HAR of daily living activities using IMU sensors in smart glasses," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng. (ECTI DAMT NCON)*, vol. 25, Mar. 2023, pp. 474–478.

[76] T. F. N. Bukht and A. Jalal, "A robust model of human activity recognition using independent component analysis and XGBoost," in *Proc. 5th Int. Conf. Advancements Comput. Sci. (ICACS)*, Lahore, Pakistan, Feb. 2024, pp. 1–7, doi: 10.1109/icacs60934.2024.10473238.

[77] S. Kamal and A. Jalal, "Multi-feature descriptors for human interaction recognition in outdoor environments," in *Proc. Int. Conf. Eng. Comput. Technol. (ICECT)*, Islamabad, Pakistan, May 2024, pp. 1–6, doi: 10.1109/icect61618.2024.10581264.

[78] P. Jantawong, N. Hnoohom, A. Jitpattanakul, and S. Mekruksavanich, "A lightweight deep learning network for sensor-based human activity recognition using IMU sensors of a low-power wearable device," in *Proc. 25th Int. Comput. Sci. Eng. Conf. (ICSEC)*, Nov. 2021, pp. 459–463.

[79] A. Alazeb, B. R. Chughtai, N. A. Mudawi, Y. AlQahtani, M. Alonazi, H. Aljuaid, A. Jalal, and H. Liu, "Remote intelligent perception system for multi-object detection," *Frontiers Neurorobotics*, vol. 18, May 2024, Art. no. 1398703, doi: 10.3389/fnbot.2024.1398703.

[80] T. F. N. Bukht, N. A. Mudawi, S. S. Alotaibi, A. Alazeb, M. Alonazi, A. A. Alarfaj, A. Jalal, and J. Kim, "A novel human interaction framework using quadratic discriminant analysis with HMM," *Comput., Mater. Continua*, vol. 77, no. 2, pp. 1557–1573, 2023, doi: 10.32604/cmc.2023.041335.

[81] N. A. Mudawi, M. Tayyab, M. W. Ahmed, and A. Jalal, "Machine learning based on body points estimation for sports event recognition," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions (ICARSC)*, Paredes de Coura, Portugal, May 2024, pp. 120–125, doi: 10.1109/icarsc61747.2024.10535954.

[82] D. Khan, A. Alshahrani, A. Almjally, N. Al Mudawi, A. Algarni, K. Alnowaiser, and A. Jalal, "Advanced IoT-based human activity recognition and localization using deep polynomial neural network," *IEEE Access*, vol. 12, pp. 94337–94353, 2024, doi: 10.1109/ACCESS.2024.3420752.

[83] M. Mushhood Afsar, S. Saqib, Y. Yasin Ghadi, S. A. Alsuhibany, A. Jalal, and J. Park, "Body worn sensors for health gaming and e-Learning in virtual reality," *Comput., Mater. Continua*, vol. 73, no. 3, pp. 4763–4777, 2022, doi: 10.32604/cmc.2022.028618.

[84] M. M. Afsar, S. Saqib, M. Aladfaj, M. H. Alatiyyah, K. Alnowaiser, H. Aljuaid, A. Jalal, and J. Park, "Body-worn sensors for recognizing physical sports activities in exergaming via deep learning model," *IEEE Access*, vol. 11, pp. 12460–12473, 2023, doi: 10.1109/ACCESS.2023.3239692.

[85] B. I. Alabdullah, H. Ansar, N. A. Mudawi, A. Alazeb, A. Alshahrani, S. S. Alotaibi, and A. Jalal, "Smart home automation-based hand gesture recognition using feature fusion and recurrent neural network," *Sensors*, vol. 23, no. 17, p. 7523, Aug. 2023, doi: 10.3390/s23177523.

[86] A. Nadeem, A. Jalal, and K. Kim, "Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness," *Symmetry*, vol. 12, no. 11, p. 1766, Oct. 2020, doi: 10.3390/sym12111766.

[87] A. A. Rafique, M. Gochoo, A. Jalal, and K. Kim, "Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13401–13430, Apr. 2023, doi: 10.1007/s11042-022-13717-y.

[88] M. Alonazi, H. Ansar, N. A. Mudawi, S. S. Alotaibi, N. A. Almujally, A. Alazeb, A. Jalal, J. Kim, and M. Min, "Smart healthcare hand gesture recognition using CNN-based detector and deep belief network," *IEEE Access*, vol. 11, pp. 84922–84933, 2023, doi: 10.1109/ACCESS.2023.3289389.

[89] M. Ullrich, A. Mücke, A. Küderle, N. Roth, T. Gladow, H. Gaßner, F. Marxreiter, J. Klucken, B. M. Eskofier, and F. Kluge, "Detection of unsupervised standardized gait tests from real-world inertial sensor data in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2103–2111, 2021.

[90] W. B. Soltana, M. Ardabilian, L. Chen, and C. B. Amar, "Adaptive feature and score level fusion strategy using genetic algorithms," in *Proc. 20th Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 4316–4319, doi: 10.1109/ICPR.2010.1049.

[91] U. Azmat, S. S. Alotaibi, M. Abdelhaq, N. Alsufyani, M. Shorfuzzaman, A. Jalal, and J. Park, "Aerial insights: Deep learning-based human action recognition in drone imagery," *IEEE Access*, vol. 11, pp. 83946–83961, 2023, doi: 10.1109/ACCESS.2023.3302353.

[92] M. Alonazi, A. M. Qureshi, S. S. Alotaibi, N. A. Almujally, N. A. Mudawi, A. Alazeb, A. Jalal, J. Kim, and M. Min, "A smart traffic control system based on pixel-labeling and SORT tracker," *IEEE Access*, vol. 11, pp. 80973–80985, 2023, doi: 10.1109/ACCESS.2023.3299488.

[93] G. De Leonardis, S. Rosati, G. Balestra, V. Agostini, E. Panero, L. Gastaldi, and M. Knaflitz, "Human activity recognition by wearable sensors: Comparison of different classifiers for real-time applications," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2018, pp. 1–6.

[94] M. Mobaraki, A. Bannadabhavi, M. J. Yedlin, and B. Gopaluni, "A vision-based deep learning platform for human motor activity recognition," in *Proc. 12th Int. Conf. Modern Circuits Syst. Technol. (MOCAST)*, Jun. 2023, pp. 1–4.

[95] Y. Abbas and A. Jalal, "Drone-based human action recognition for surveillance: A multi-feature approach," in *Proc. Int. Conf. Eng. Comput. Technol. (ICECT)*, Islamabad, Pakistan, May 2024, pp. 1–6, doi: 10.1109/icect61618.2024.10581378.

[96] H. Ansar, A. Ksibi, A. Jalal, M. Shorfuzzaman, A. Alsufyani, S. A. Alsuhibany, and J. Park, "Dynamic hand gesture recognition for smart lifecare routines via K-ary tree hashing classifier," *Appl. Sci.*, vol. 12, no. 13, p. 6481, Jun. 2022, doi: 10.3390/app12136481.

[97] B. Debnath, M. O'Brien, S. Kumar, and A. Behera, "Attentional learnable pooling for human activity recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13049–13055.

[98] D. Khan, N. A. Mudawi, M. Abdelhaq, A. Alazeb, S. S. Alotaibi, A. Algarni, and A. Jalal, "A wearable inertial sensor approach for locomotion and localization recognition on physical activity," *Sensors*, vol. 24, no. 3, p. 735, Jan. 2024, doi: 10.3390/s24030735.

[99] N. A. Mudawi, M. Pervaiz, B. I. Alabduallah, A. Alazeb, A. Alshahrani, S. S. Alotaibi, and A. Jalal, "Predictive analytics for sustainable E-learning: Tracking student behaviors," *Sustainability*, vol. 15, no. 20, p. 14780, Oct. 2023, doi: 10.3390/su152014780.

[100] N. Al Mudawi, H. Ansar, A. Alazeb, H. Aljuaid, Y. AlQahtani, A. Algarni, A. Jalal, and H. Liu, "Innovative healthcare solutions: Robust hand gesture recognition of daily life routines using 1D CNN," *Frontiers Bioeng. Biotechnol.*, vol. 12, Jul. 2024, Art. no. 1401803, doi: 10.3389/fbioe.2024.1401803.

[101] S. J. Hashmi, B. Alabdullah, N. A. Mudawi, A. Algarni, A. Jalal, and H. Liu, "Enhanced data mining and visualization of sensory-graph-modeled datasets through summarization," *Sensors*, vol. 24, no. 14, p. 4554, Jul. 2024, doi: 10.3390/s24144554.

[102] N. A. Almujally, D. Khan, N. A. Mudawi, M. Alonazi, A. Alazeb, A. Algarni, A. Jalal, and H. Liu, "Biosensor-driven IoT wearables for accurate body motion tracking and localization," *Sensors*, vol. 24, no. 10, p. 3032, May 2024, doi: 10.3390/s24103032.

[103] A. Raza, S. A. Chelloug, M. H. Alatiyyah, A. Jalal, and J. Park, "Multiple pedestrian detection and tracking in night vision surveillance systems," *Comput., Mater. Continua*, vol. 75, no. 2, pp. 3275–3289, 2023, doi: 10.32604/cmc.2023.029719.

[104] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene semantic recognition based on modified fuzzy C-mean and maximum entropy using object-to-object relations," *IEEE Access*, vol. 9, pp. 27758–27772, 2021, doi: 10.1109/ACCESS.2021.3058986.

[105] M. Muneeb, H. Rustam, and A. Jalal, "Automate appliances via gestures recognition for elderly living assistance," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Lahore, Pakistan, Feb. 2023, pp. 1–6, doi: 10.1109/ICACS55311.2023.10089778.

[106] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1010–1019.

[107] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," 2019, *arXiv:1905.04757*.

[108] W. Cao, J. Zhong, G. Cao, and Z. He, "Physiological function assessment based on Kinect V2," *IEEE Access*, vol. 7, pp. 105638–105651, 2019.

[109] R. Cohen, G. Fernie, and A. R. Fekr, "Contactless drink intake monitoring using depth data," *IEEE Access*, vol. 11, pp. 12218–12225, 2023.

[110] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Trans. Image Process.*, vol. 29, pp. 3835–3844, 2020.

MOUAZMA BATOOL is currently pursuing the Ph.D. degree with Air University, Pakistan. Her research interests include wearable and optical sensors, signal acquisition, the IoT, and life-log generation.

MONEERAH ALOTAIBI, photograph and biography not available at the time of publication.

SULTAN REFA ALOTAIBI, photograph and biography not available at the time of publication.

DINA ABDULAZIZ ALHAMMADI is currently an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, specializing in artificial intelligence. Her research interests include AI, human–computer interaction, personality recognition, and social AI. She is actively involved in mentoring students and collaborating with industry partners to bridge the gap between academia and practical applications of AI.

MUHAMMAD ASIF JAMAL received the bachelor's degree in computer engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2009, and the master's degree in computer engineering from Mid Sweden University, Sundsvall, Sweden, in 2014. He is currently on study leave from his position as a Lecturer with the Department of Computer Science, Air University, Islamabad, Pakistan, while pursuing the Ph.D. research in medical image processing with the MIP Laboratory, Chosun University, South Korea. His research interests include medical image segmentation and classification and wireless sensor networks.

AHMAD JALAL received the Ph.D. degree from the Department of Biomedical Engineering, Kyung Hee University, Republic of Korea. He is currently an Associate Professor with the Department of Computer Science and Engineering, Air University, Pakistan. He is also a Postdoctoral Research Fellowship with POSTECH. His research interests include multimedia contents and artificial intelligence.

BUMSHIK LEE received the B.S. degree from Korea University, Seoul, South Korea, and the M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was a Research Professor with KAIST, in 2014, and a Postdoctoral Scholar with the University of California at San Diego (UCSD), CA, USA, from 2012 to 2013. He was a Principal Engineer with the Advanced Standard R&D Laboratory, LG Electronics, Seoul. He is currently a Professor with the Department of Information and Communications Engineering, Chosun University, South Korea. His research interests include video compression, image and video processing, video security, and medical image processing.