IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 February 2024, accepted 26 February 2024, date of publication 7 March 2024, date of current version 13 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373199

TOPICAL REVIEW

Human Action Recognition Systems: A Review of the Trends and State-of-the-Art

MISHA KARIM[1], SHAH KHALID[1], ALIYA ALERYANI[2], JAWAD KHAN[3,4], IRFAN ULLAH[5], AND ZAFAR ALI[6]

[1]Department of Computing, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[3]Department of AI Software, Gachon University, Seongnam-si 13120, South Korea
[4]Department of Robotics, Hanyang University, Ansan 15588, South Korea
[5]Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal 18050, Pakistan
[6]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

Corresponding author: Shah Khalid (shah.khalid@seecs.edu.pk)

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Small Groups RGP.1/369/44.

ABSTRACT Human action recognition (HAR), deeply rooted in computer vision, video surveillance, automated observation, and human-computer interaction (HCI), enables precise identification of human actions. Numerous research groups have dedicated their efforts to various applications and problem domains in HAR systems. They trained classification models using diverse datasets, enhanced hardware capabilities and employed different metrics to assess performance. Although several surveys and review articles have been published periodically to highlight research advancements in HAR, there is currently no comprehensive and up-to-date study that encompasses architecture, application areas, techniques/algorithms, and evaluation methods as well as challenges and issues. To bridge this gap in the literature, this article presents a comprehensive analysis of the current state of HAR systems by thoroughly examining a meticulously chosen collection of 135 publications published within the past two decades. These findings have implications for researchers engaged in different aspects of HAR systems.

INDEX TERMS Human action recognition, computer vision, machine learning, video surveillance, HAR architecture.

I. INTRODUCTION

Human action recognition (Human Action Recognition (HAR)), a crucial component of computer vision, lies at the intersection of the latest sensor technologies, Machine Learning (ML) and Deep Learning (DL). This intricate process involves careful identification and classification of human actions based on raw data collected from diverse sources, such as body-worn sensors, smartphones, and cameras. By analyzing the patterns of human movement, these actions enable a deeper understanding of behavior across various environments, offering valuable insights for applications ranging from healthcare to law enforcement [1].

The realm of HAR encompasses a wide range of practical applications, including the enhancement of surveillance systems for crime prevention and the improvement of user interactions in smart home settings. In recent years, this field has undergone significant advancements with HAR systems evolving from basic motion sensors to sophisticated networks capable of intricate pattern recognition and real-time processing [2], [3], [4]. Figure 1 demonstrates the utilization of deep learning (DL) techniques on sensory data for time-series categorization, which plays a vital role in understanding temporal changes and identifying human actions using various sensor modalities [5], [6].

HAR has made significant progress, but still faces considerable challenges. Throughout its development, the field has achieved important milestones, encountered obstacles, and adapted to emerging technologies. Initially, the main challenge was to achieve high accuracy in various complex real-world scenarios. This requires a deep understanding of the intricacies of human movement and real-time data

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

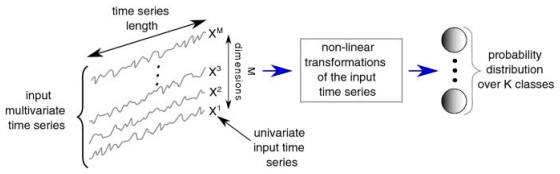

FIGURE 1. DL neural networks used for classifying the time series stamp [3].

processing. Researchers have addressed this challenge by creating inclusive datasets covering a wide range of human activities and demographics. They also focused on reducing the biases in algorithmic models and utilizing multimodal data to enhance the robustness and accuracy of HAR systems [5], [6].

Furthermore, integrating HAR with emerging technologies, such as augmented reality (AR) and virtual reality (VR), has opened up new possibilities, providing more immersive and interactive experiences. This integration highlights the dynamic nature of the field because it adapts to the potential offered by these technologies [7], [8]. In this review, our objective was to offer a comprehensive overview of the development of HAR. We discuss key milestones, technological components that have supported its evolution, and recent advancements. Additionally, we explore the essential features necessary for accurate recognition, address the ongoing challenges that shape the future of HAR research, and critically evaluate current approaches.

- A thorough investigation of the most recent advancements in HAR, encompassing methodologies based on ML or DL, benchmark datasets, and the metrics employed to assess their effectiveness.
- A detailed presentation of a standard HAR system's architecture, outlining the crucial technological components and their interaction.
- A critical evaluation of current approaches in HAR using standard metrics.

The remainder of this paper is structured as follows: Section II provides an in-depth explanation of the methodology employed for the literature search and selection, Section III summarizes the main findings, and Section IV presents concluding remarks and potential avenues for future research.

II. MATERIALS AND METHODS

This section outlines the methodological framework used to conduct a systematic review of the literature on HAR systems. The primary objective of this systematic approach is to ensure comprehensive coverage and robust evaluation of the relevant research published within the scope of this study.

The selection process is depicted in Figure 2, which adheres to PRISMA guidelines.

A. SEARCH STRATEGY

To comprehensively gather studies related to HAR based on sensor and vision technologies, a thorough search strategy was implemented. This strategy involved systematically exploring various electronic databases, such as Google Scholar, Web of Science, Scopus, IEEE Xplore, and ScienceDirect. By utilizing a combination of specific keywords like "human action recognition," "daily life activities," "sensor data," and "vision data," along with Boolean operators (AND, OR, NOT), the search was carefully refined to obtain relevant results. The search strings used for each database can be found in Table 1.

B. INCLUSION AND EXCLUSION CRITERIA

The inclusion criteria were precisely defined to select studies that significantly contributed to the understanding of HAR systems. Criteria for exclusion were established to uphold the quality of the review, ensuring a focus on original empirical research.

Inclusion criteria included:

1) Studies published in English from 2000 to 2023.
2) Research focusing on HAR within the context of the Activities of Daily Living (ADL).
3) Studies based on sensors or vision data.
4) Articles published in peer-reviewed journals or conference proceedings.
5) Studies with accessible full text.

Exclusion criteria applied:

1) Non-English publications.
2) Duplicate studies within the databases.
3) Review articles and meta-analyses, which were used for background context.
4) Studies with inaccessible full text or published in low-quality venues.

C. SELECTION PROCESS

A total of 164 articles were obtained from the initial database searches, comprising 160 records from the database searches and an additional 4 records sourced from other relevant sources. After eliminating duplicates, 128 distinct records were retained. These records were screened based on their titles and abstracts to determine their relevance.

D. DATA EXTRACTION AND ANALYSIS

To maintain consistency and objectivity throughout the review process, a standardized form was used for data extraction. This form encompasses various details, such as the authors, publication year, methodologies employed, datasets used, evaluation metrics, and principal findings. Subsequently, a mixed-methods approach was employed to synthesize the extracted data, enabling the identification of prevailing trends, state-of-the-art techniques, and persistent challenges within the domain of HAR research.

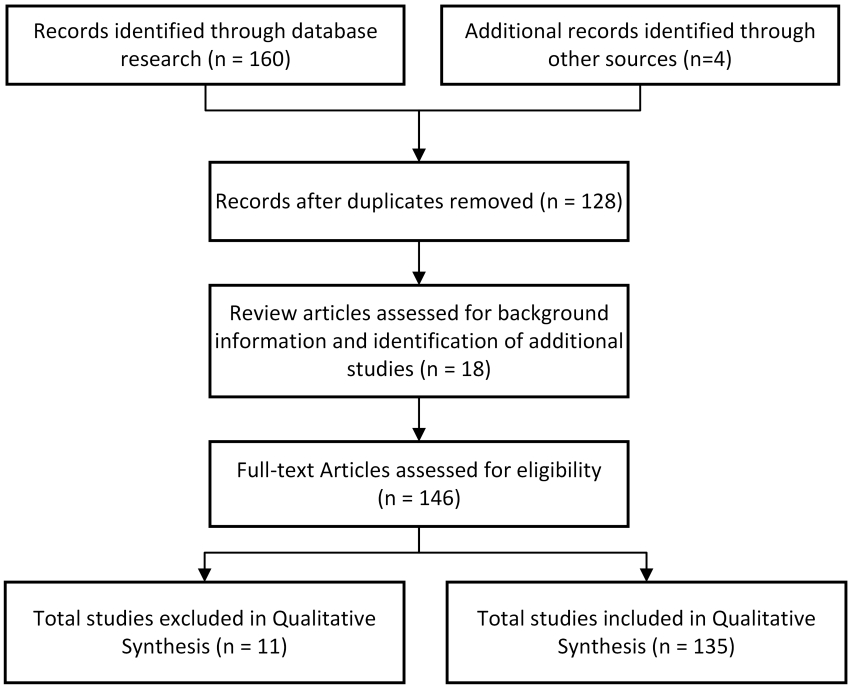

FIGURE 2. The PRISMA flowchart illustrates the meticulous selection process, starting from the initial identification of records to the ultimate inclusion for qualitative synthesis.

TABLE 1. Search strings used in searching each database.

Database	Search String
Google Scholar	("Human Action Recognition" OR "HAR") AND ("Daily Life Activities" OR "DLA") AND ("Sensor Data" OR "Vision Data")
Web of Science	("Human-Computer Interaction" OR "HCI") AND ("Multimodal Data Fusion") AND ("Sensor Networks" OR "Vision Data")
Scopus	("Activity Recognition" OR "AR") AND ("Wearable Sensors" OR "Wearable Devices") AND ("Daily Activities" OR "DLA") AND ("Deep Learning" OR "Machine Learning")
IEEE Xplore	("Action Recognition") AND ("Computer Vision") AND ("Daily Activities" OR "DLA") AND ("Healthcare" OR "Rehabilitation")
ScienceDirect	("Transfer Learning" OR "Domain Adaptation") AND ("Sensor Data" OR "Vision Data") AND ("HAR" OR "Activity Recognition")

E. RATIONALE FOR INCLUSION AND EXCLUSION CRITERIA

The inclusion and exclusion criteria were established to focus the review on methodologically robust studies that directly pertained to the sensor- and vision-based HAR systems. This deliberate selection aimed to ensure the academic rigor and relevance of the review.

III. SUMMARY OF KEY OBSERVATIONS

This section provides an overview of the key findings from the present study, which include the structure of HAR systems and their applications, datasets used, algorithms employed, and the challenges faced during the research process.

A. THE HAR SYSTEM ARCHITECTURE

HAR systems have been developed to analyze raw sensory data and derive meaningful insights regarding human behavior. The initial step involves collecting data from various sensors, including cameras, inertial measurement units, and microphones, which capture multi-modal data reflecting human movement. To ensure data quality and standardization, the collected data undergoes preprocessing, which includes labeling the data with relevant action descriptors and normalizing the input scales. These preprocessing steps are essential for cleaning and preparing the data for feature representation. Subsequently, sophisticated techniques are employed to extract significant features from the preprocessed data. In the classification stage, advanced DL or artificial intelligence (AI) models interpret these features to categorize complex actions. Convolutional Neural Networks (Convolutional Neural Network (CNN)s) and Recurrent Neural Networks (RNNs) are commonly utilized algorithms for discerning and classifying actions.

This architectural framework serves as a fundamental basis for comprehending the intricate challenges associated with HAR. These challenges encompass the handling of high-dimensional data, ensuring reliable feature extraction in

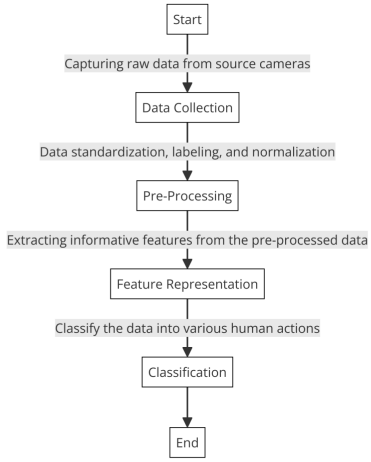

FIGURE 3. The process of a HAR system is depicted in this flowchart. It visually represents the various interconnected stages that are crucial for identifying human actions. Every step, starting from data collection and ending with classification, plays a vital role and relies on each other, ultimately enhancing the accuracy and effectiveness of the entire system.

varying environmental conditions, and selecting appropriate classification algorithms that offer both accuracy and prompt response times. The subsequent sections of this study delve into these challenges in greater detail, proposing potential solutions and discussing diverse applications and the range of techniques employed to refine HAR systems.

Figure 4 provides a taxonomy that encapsulates the existing literature on HAR, illustrating the extensive scope and depth of the field. The following sections will further explore various applications of HAR, showcasing the wide array of techniques and algorithms that enhance the capabilities of these systems.

B. APPLICATION AREAS

This section provides an overview of the various fields where HAR finds its applications, with particular emphasis on its important contributions to ADL, surveillance, Human-Computer Interaction (HCI), and competitive sports.

1) HAR IN DAILY LIFE ACTIONS

To enhance human well-being, a nuanced understanding of daily activities facilitated by advanced HAR technologies such as PoseNet and GHUM is crucial. These models, which utilize convolutional neural networks, have significantly advanced human pose estimation in images and videos, leading to breakthroughs in fields such as smart home systems, elderly care, and physiotherapy [9], [10], [11], [12].

PoseNet's efficiency in detecting and tracking angular movements was highlighted by its impressive accuracy rate of 97.6% for 2D pose detection. This capability has been instrumental in telehealth systems for home-based rehabilitation, allowing for the precise monitoring and adjustment of rehabilitation exercises [10]. GHUM, with its advanced bone-based sensing technique, offers robust pose estimation in nonintrusive monitoring scenarios, making it particularly useful in elderly care environments for fall detection and activity monitoring [13]. These technologies ensure safety and health monitoring in smart homes by providing real-time alerts in case of abnormalities [14].

Despite these advancements, HAR systems face challenges related to their sensitivity to varying lighting conditions and camera angles. These environmental factors can significantly affect the accuracy of pose estimation, with deviations of up to 15% caused by variations in the lighting conditions [15]. To address these issues and enhance the robustness of HAR systems, multiview models and image enhancement techniques have been developed [16]. Additionally, recognizing individual actions in complex scenarios involving multiple people poses challenges. Recent studies explored advanced detection models to address these complexities [11].

Gong et al. [17] presented a novel DL model that effectively improved the recognition accuracy of activity classes within datasets, particularly in complex activity scenarios. By incorporating this model, recognition accuracy was enhanced by approximately 10% [17], [18]. This significant development addressed the persistent challenge of the semantic gap in HAR.

Ongoing research endeavors are expected to prioritize the enhancement of data quality, model adaptability, and real-time processing in HAR technologies. These advancements are of utmost importance for accurately interpreting intricate daily human actions and seamlessly integrating HAR into our everyday lives. Ultimately, these advancements have the potential to revolutionize interactions with technology.

2) HAR IN HCI

The integration of HAR into HCI has brought about a significant transformation in how users interact with technology. By integrating HAR capabilities, such as gesture control and eye tracking, systems have been able to enhance the user experience by making interactions more intuitive and responsive. These technologies have become increasingly prevalent in various applications, ranging from smart home interfaces to healthcare monitoring, and have significantly improved the seamless interaction between humans and technology [19], [20].

One notable example of the impact of HAR technologies is on smart home systems, where they have achieved an impressive accuracy rate of up to 90% in gesture recognition. This high level of accuracy enables users to control home appliances more naturally and efficiently [21]. In the context of healthcare monitoring, HAR systems play a crucial role in facilitating patient interaction with medical devices. By doing so, they contributed to the ease and accuracy of patient data collection.

Overall, integrating HAR into HCI has revolutionized user interactions with technology, making them more intuitive, responsive, and seamless. These advancements have been observed in various domains, including research showing that gesture recognition accuracy can decrease by up to 20%

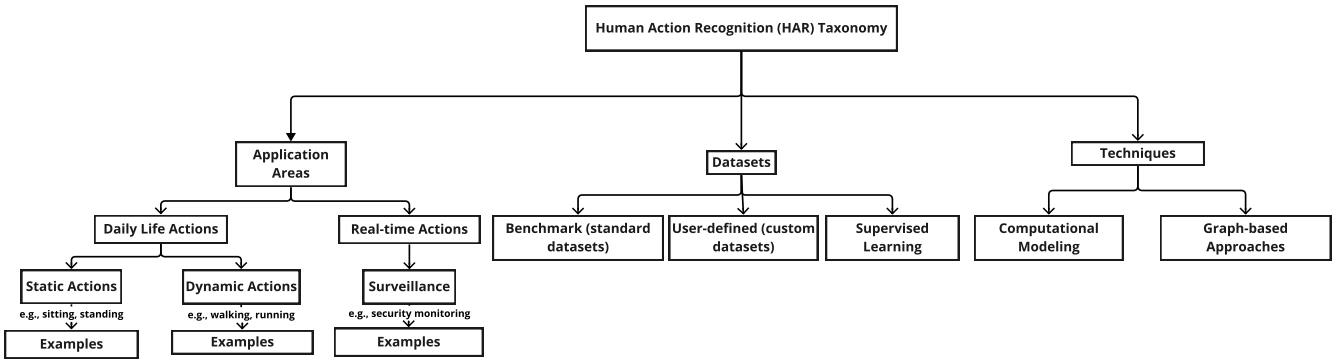

FIGURE 4. Taxonomy of HAR. This taxonomy outlines the various facets of HAR research, encompassing methodologies, applications, and challenges, providing a structured overview of the field.

in poorly lit environments [22]. Furthermore, the latency in eye-tracking systems, which is crucial for interactive applications, can vary, with some systems reporting latencies of up to 50 ms, which may affect user experience in fast-paced applications [23].

This study aims to overcome these limitations by developing robust HAR algorithms. These algorithms have been specifically designed to adapt to different environmental conditions and user behaviors, thereby enhancing the accuracy and minimizing the latency. For instance, the use of machine learning (ML) techniques in dynamic mathematics software has resulted in a 15% enhancement in user interaction efficiency by predicting user intent [24].

In the future, the field of HAR for HCI is expected to witness further advancements with the integration of AR/VR technologies. These advancements are geared towards creating more immersive and interactive experiences, potentially reducing the perceived latency in user interactions and enhancing the naturalness of user interface controls. The integration of HAR with AR/VR holds the promise of opening up new possibilities for HCI, fostering greater user engagement, and facilitating the development of innovative interaction paradigms.

3) HAR IN COMPETITIVE SPORTS

The integration of HAR into competitive sports has brought about significant changes, particularly in the utilization of biomechanical data to enhance athletic performance and develop strategies for preventing injuries. Sports analytics relies heavily on technologies such as accelerometers and GPS tracking devices, which provide detailed insights into athlete movements and training loads. These insights are crucial for personalizing training programs and minimizing the risk of injury [15], [25], [26].

For example, in the track and field, HAR technologies have been employed to optimize the running techniques. The data showed that athletes could achieve up to a 5% improvement in running efficiency by adjusting their form based on HAR feedback [27]. Similarly, in team sports, HAR systems have played a vital role in reducing injury rates by approximately 20% through tailored training programs and the early detection of fatigue [28], [29].

Despite the achievements of HAR, several challenges need to be addressed. One of these challenges is the variable accuracy of HAR systems under various environmental conditions. For instance, in outdoor sports, data collection can be hindered by weather variations, resulting in a decrease in data reliability of up to 10% under adverse weather conditions [30], [31], [32]. Moreover, the complexity of athlete movements and the presence of sports equipment can affect accuracy.

In the future, advancements in HAR for sports will primarily focus on improving the precision of collected data and developing more sophisticated analytics. The integration of HAR with ML, DL, and AI has great potential for transforming athletic performance analysis into predictive analytics. By utilizing HAR data for real-time feedback and for predicting injury risks, more effective training strategies and injury prevention measures can be implemented [33]. The evolving role of HAR in competitive sports is expected to provide deeper insights into athletic performance and become an integral aspect of sports training and performance enhancement.

4) HAR IN SURVEILLANCE

The significance of HAR in surveillance, particularly in security and elder care, has grown significantly. HAR systems play a crucial role in improving the efficiency and privacy of real-time action identification in surveillance operations, thereby contributing to monitoring environments and detecting unusual activities [34], [35]. In the context of elderly care facilities, HAR systems have proven to be effective in reducing false-alarm rates by up to 30%, leading to a notable improvement in emergency response effectiveness [36].

The future of HAR in surveillance is characterized by the integration of AI tools, which are expected to enable the development of more context-aware and intelligent

systems. This integration aims to enhance the ability of the system to accurately interpret complex scenarios, with recent advancements demonstrating a 25% improvement in detecting subtle activities [37]. Such advancements are particularly crucial in domains such as elderly care, where accurately distinguishing normal behaviors from potential emergencies can significantly impact the safety and well-being of individuals.

The application of HAR technology in different settings poses a significant challenge. In particular, outdoor security scenarios often present difficulties for surveillance systems owing to unpredictable weather conditions and varying lighting. These factors can have a detrimental effect on the accuracy of activity detection, reducing it by as much as 20% [38], [39]. It is crucial to ensure the reliability and precision of HAR systems under diverse conditions, particularly for applications in security and healthcare. Ongoing research endeavors are focused on enhancing the robustness of HAR systems under such diverse conditions, such as low light and occlusions, especially for applications in security and healthcare. This involves the development of advanced algorithms and the integration of ML/DL techniques to improve adaptability to changing conditions and camera perspectives. As HAR technologies continue to progress, their integration with AI is expected to result in more sophisticated surveillance systems. This integration enhances the efficiency of both action detection and contextual interpretation, aligning surveillance operations more closely with real-world requirements [32], [40], [41].

C. DATASETS

The HAR field relies heavily on a wide range of datasets, each providing unique insights and presenting fundamental challenges that are crucial for advancing HAR technologies. The datasets used in HAR include both RGB and depth data, which are distinct modalities. While Red Green Blue (RGB) images capture color information, depth images are two-dimensional and utilize Time-of-Flight (TOF) technology. The RGB format merges both color and depth data and displays the distances of the objects in the RGB images from the image plane. These datasets are pivotal for developing and testing HAR algorithms and significantly contribute to the progress of the field [42].

1) BENCHMARK DATASETS

Benchmark datasets play a vital role in (HAR) research because they serve as a fundamental foundation for the development and assessment of models. These datasets hold immense importance, as they provide controlled settings for testing and enhancing the HAR algorithms.

KTH dataset introduced by Schuldt et al. [43], which comprises six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed by 25 subjects in four different scenarios, consists of 2391 video sequences with a resolution of 160 × 120 pixels and a frame rate of 25 fps. This dataset has certain advantages such as being publicly available for non-commercial use, having a large number of samples, and covering various action categories and scene variations. However, it also has some disadvantages, such as low resolution, homogeneous background, and the absence of complex interactions or occlusions. The KTH dataset can be utilized to test and compare different methods for action recognition and detection, such as grid key points, video generation, and temporal attention units. Future development recommendations include enhancing the resolution and diversity of videos, incorporating more challenging actions and interactions, and integrating additional modalities, such as depth or skeleton, as noted by Baccouche et al. [44].

UT-Kinect: This dataset is presented by Xia et al. [45], and is commonly used for action recognition based on depth sequences. This dataset was acquired using a single stationary Kinect camera and featured 10 action types: walking, sitting down, standing up, picking up, carrying, throwing, pushing, pulling, waving hands, and clapping hands. Each action type was performed by ten different subjects, resulting in a total of 100 samples. The dataset included three channels of information: RGB, depth, and skeleton joint location. A high resolution of 640 × 480 pixels and fixed frame rate of 30 fps are among the advantages of this dataset. However, it has a limited number of samples and action categories and includes no variations or occlusions. Cai et al. [46] have effectively utilized its data, but its controlled conditions might not fully capture real-world intricacies. Future developments of this dataset could include increasing the number and diversity of samples, adding more action categories and scene variations, and incorporating additional sensors or viewpoints.

MSR 3D Action dataset comprises twenty distinct action categories and ten subjects, and each subject performed each action two or three times. A total of 567 depth map sequences were recorded, with a resolution of 320 × 240 pixels. The data were captured using a depth sensor similar to the Kinect device. Some of the advantages of this dataset are that it is publicly available for non-commercial use, and covers a wide range of action categories including daily, health-related, explored by Xia et al. [45] and Ben Tamou et al. [47], the dataset has been utilized in various scenarios, including testing and comparing different methods for action recognition from depth maps such as Signed Distance Function (SDF), Subspace Video Linear Regression Model (SVLRM), and Dynamic Kernel Network (DKN). For future development, it is recommended to increase the resolution and diversity of the depth maps, incorporate additional modalities such as RGB or skeletons, and include more scene variations or occlusions.

Florence 3D Action: This dataset comprises nine everyday activities: wave, drink from a bottle, answer the phone, clap, tie lace, sit down, stand up, read/watch, and bow. The dataset features ten subjects, each of whom performs the actions two or three times. The dataset included 215 video clips with a resolution of 480 × 640 pixels and frame rate of 25 fps,

TABLE 2. Summary of benchmark datasets - part 1.

Dataset	Modality	Frame Rate	Resolution
KTH (2D-3D)	RGB, Depth	0.066-25 fps	160x120
UT-Kinect (3D)	RGB, Depth, Skeleton joint locations	30 fps	480x640 for RGB, 320x240 for Depth
MSR-Action-3D	Depth, Skeleton joint locations	15 fps	320x240
Florence-Action	Depth, Skeleton joint locations	25 fps	1280x720 for HD video
NTU RGBD 3D	RGB, Depth, 3D Skeleton, Infrared	30 fps	1920x1080 (depth maps), 512x424 (infrared sequences)
HuDa-3DAct	RGB, Depth, 3D Skeleton, Infrared	15 fps	640x480
UCF-101	RGB	25 fps	320x240
WISDM	Accelerometer, Gyroscope	20 Hz	Not specified.
UCI HAR	Accelerometer, Gyroscope	50 Hz	Not specified.

captured using a Kinect camera. The dataset is characterized by its provision of RGB, depth, and skeleton information as well as its high resolution and coverage of common actions in daily life [48]. The dataset was used to test and compare various methods of action recognition using depth cameras, such as structured keypoint pooling, STM, and R2+1D-BERT. Recommendations for future development include increasing the number and diversity of samples, adding more action categories and scene variations, and incorporating more sensors and perspectives.

NTU RGB D: As utilized by Tu et al. [49], this large-scale dataset is a comprehensive resource for RGB HAR, encompassing 56,880 samples from 60 action classes across 40 subjects. The action categories included 40 daily actions, nine health-related actions, and 11 mutual actions, occurring in 17 different scene conditions captured by three cameras at various horizontal imaging viewpoints. By providing multimodal information, including depth maps, 3D skeleton joint positions, RGB frames, and infrared sequences, the dataset is publicly available for non-commercial use. Although advantageous for its extensive samples, action categories, and scene conditions, it suffers from low resolution (320 $\times$ 240 pixels), noise, missing values in depth and skeleton data, and a lack of complex interactions or occlusions. Usage scenarios involve testing and comparing action recognition and detection methods using RGB data such as PoseC3D, VideoMAE, and OTI. Future development recommendations include improving the RGB data resolution and quality, introducing complex interactions and occlusions, and incorporating additional sensors and viewpoints.

RGB D Huda Act: Explored by Zhao et al. [50], it is a comprehensive collection of RGB HAR data. It encompasses 56,880 samples from 60 action classes, collected from 40 subjects. These actions can be broadly categorized into three groups: 40 daily actions (e.g., drinking, eating, reading), nine health-related actions (e.g., sneezing, staggering, falling), and 11 mutual actions (e.g., punching, kicking, hugging). The actions were captured under 17 different scene conditions corresponding to 17 video sequences (S001-S017). The actions were recorded using three cameras with varying horizontal imaging viewpoints ($-45°$, $0°$, and $+45°$). The dataset provided multimodal information for action characterization, including depth maps, 3D skeleton joint positions, RGB frames, and infrared sequences. The NTU RGB D dataset offers several advantages, such as public availability for noncommercial use, a large number of samples, action categories, scene conditions, and coverage of various action types and viewpoints. However, it also has some limitations, including low resolution (320 $\times$ 240 pixels), noise, missing values in the depth and skeleton data, and a lack of complex interactions or occlusions. The dataset can be utilized to test and compare different methods for action recognition and detection using RGB data such as PoseC3D, VideoMAE, and OTI. Recommendations for future development include improving the resolution and quality of RGB data, incorporating more complex interactions and occlusions, and introducing additional sensors and viewpoints.

UCF 101: Examined by He et al. [51], it comprises 101 categories such as sports and human-object interactions. Using 13,320 clips sourced from YouTube, the dataset presents challenges with variations in the camera motion, object appearance, and environmental conditions. The advantages include diverse collections and robust model performance. The limitations include limited resolution and potential noise in the action labels. Suited for applications such as surveillance, it may not be ideal to recognize simple actions using other modalities. Recommendations for future development include enhancing the RGB video resolution, refining action labels, and incorporating additional modalities, such as depth or skeleton, for improved recognition.

Each dataset has distinct characteristics that make it suitable for different aspects of HAR research. However, limitations such as realism, diversity, and complexity should be considered when selecting a dataset for a specific research goal.

2) USER-GENERATED DATASETS

In computer vision research, particularly in applications such as surveillance systems, home monitoring [52], [53], [54], [55], [56], [57], [58] and sensor-based applications for senior monitoring [59], HAR is among key areas of interest. The role of human motion-based characteristics in detection and classification, involving pre-processing techniques such as spatial-temporal filtering, background subtraction, and optical flow [60], along with applications

in video surveillance [61] and virtual reality [62], often leverages both benchmark and user-generated datasets.

a: VIHASI

Ragheb et al. [63] developed the ViHASi dataset, a large collection of synthetic human activities designed for testing action identification algorithms. This dataset included synchronized perspectives from multiple cameras, various actors, and a wide range of action classes, thereby offering a comprehensive testing ground for long video sequences with numerous action samples.

b: SMARTPHONE-BASED DATASETS

Micucci et al. [64] examined smartphone accelerometer datasets for detecting ADLs. They highlighted the importance of feature selection in categorizing falls, using the UniMiB SHAR dataset as a case study. However, datasets such as MobiAct, RealWorld, and UMA Fall demonstrate a gender imbalance, predominantly featuring male subjects, which should be considered in future research.

c: VIDOR

Shang et al. [65] introduced the VidOR dataset containing annotated films with object categories and predicates. Their work included an automated pipeline for labeling user-submitted videos and extensive annotation analysis, offering rich resources for video object and relation recognition.

Table 4 offers a more comprehensive perspective on the performance of classifiers in terms of accuracy, considering both benchmark and user-generated datasets.

3) SUMMARY ON COMPARING HAR DATASETS

In summary, notable advancements have been achieved in creating HAR datasets, contributing significantly to the progress of this field. A thorough examination of different datasets demonstrated substantial improvements in addressing complex human actions and interactions. However, challenges still need to be addressed, particularly in terms of dataset diversity and representativeness, as these factors can impact the universality and effectiveness of HAR systems. It is crucial to obtain datasets that encompass a wide range of human actions under various conditions. Future developments should focus on enhancing the comprehensiveness of datasets by prioritizing inclusivity and variability to better reflect real-world scenarios. This approach will further refine HAR technologies, resulting in more precise and adaptable systems, ultimately pushing the boundaries of what can be achieved in human-action recognition.

D. TECHNIQUES/ALGORITHMS

Technological progress has had a profound influence in different areas, particularly in the development of innovative techniques in image detection, computer vision, and facial recognition, which have played a vital role in advancing HAR with applications in training, security, video surveillance, and

TABLE 3. Summary of benchmark datasets - part 2.

Dataset	Samples/Classes	Application Scenarios
KTH (2D-3D)	2 versions with 6 actions, 25 subjects in 4 scenarios	Basic action recognition, motion analysis
UT-Kinect (3D)	10 subjects, 10 actions, 3D joint locations	Kinect-based interaction, gesture recognition
MSR-Action-3D	10 subjects, 20 actions, 3D joint locations	Gaming, virtual reality, advanced gesture recognition
Florence-Action	10 subjects, 9 actions, 3D joint locations	Daily activity monitoring, elder care
NTU RGBD-3D	60 actions, 40 subjects, captured by 3 cameras	Comprehensive activity analysis, multi-person interaction
HuDa-3DAct	12 ASL gestures, 10 subjects, segmented hand frames	Sign language recognition, detailed hand gesture analysis
UCF-101	13,320 samples, 101 action classes	Broad range of actions, video surveillance, sports analysis
WISDM	18 activities, 51 subjects, smartphone and smartwatch data	Wearable device applications, health monitoring
UCI HAR	Six activities, 30 subjects, smartphone data, 561-feature vector	Mobile health applications, fitness tracking

TABLE 4. A broader view of classifiers' performance on some benchmark and user-generated datasets.

Author	Dataset	Tool/Framework	Classifier	Accuracy	Scope of Use
[66]	WISDM	WEKA and Adaboost	Decision Stump, Random Tree, RF, Hoeffding Tree, REP Tree	97.83%	Activity recognition from smartphone sensors
[67]	UCI HAR	Score level and feature level fusion	K-NN, SVM	97.12%	Human activity recognition using smartphone accelerometer and gyroscope data
[68]	UCF-101	ARTNet	ARTNet	94.3%	Action recognition in UCF-101 video dataset
[69]	NTU RGB D	MATLAB, STNN	CNN	96.3%	Action recognition from 3D skeletal data in NTU RGBD dataset
[70]	WISM Smartphone Activity	WEKA, RF, Bagging	RF	87.19%	Smartphone activity recognition for personalized applications

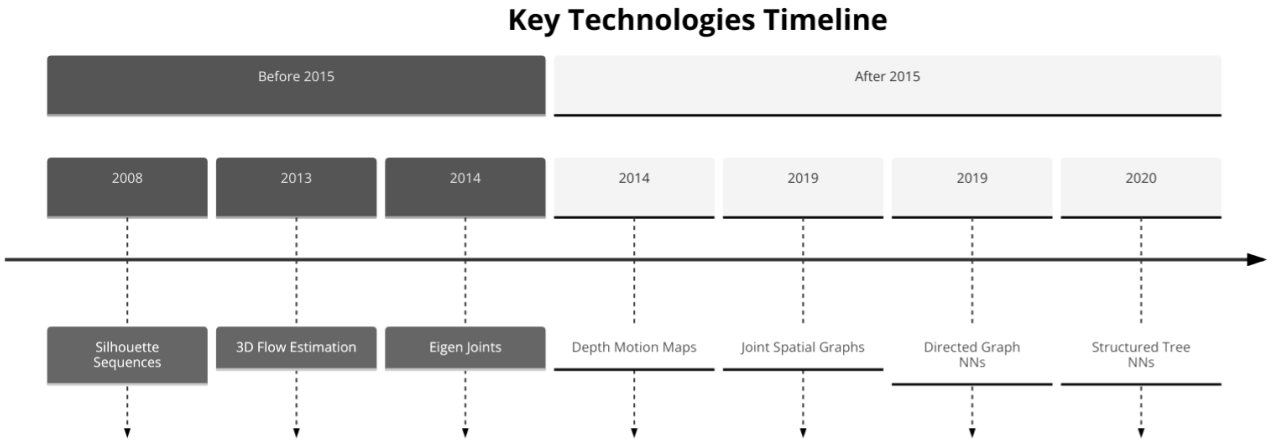

FIGURE 5. The evolution of HAR strategies (from 2008 to 2020).

automated observation. Figure 5 illustrates the progression of the HAR strategies over time, emphasizing important milestones in the field.

1) SUPERVISED LEARNING

Supervised learning methods have played a crucial role in advancing HAR, as evidenced by numerous studies that show their effectiveness.

Random Forest (RF): Xu et al. [71] demonstrated the efficacy of Random Forest (RF) in handling accelerometer data by achieving superior accuracy and adaptability to environmental limitations when compared to conventional approaches.

DL with Wearable Sensors: employed ML algorithms in conjunction with wearable sensors found in smartwatches. Their study highlights the practical implications of these technologies in the realm of health and safety monitoring [72].

3-D CNN for Video Surveillance: Almaadeed et al. [73] presented a groundbreaking Three-Dimensional Convolutional Neural Network (3D CNN) structure that effectively enhances the analysis of action sequences and strengthens the capabilities of video surveillance.

Structured-Tree Neural Networks: In their study, Khan et al. [69] introduced an innovative structured-tree neural network that was trained on the NTU RGB D dataset. This approach provides a distinct perspective for analyzing human motion in the context of 3D action recognition.

Enhancement of Pre-trained CNNs: Ozcan and Basturk [74] investigated the performance enhancement of the NASNet-Large architecture in pre-trained CNNs. They conducted experiments to demonstrate the improved accuracy achieved by fine-tuning training parameters.

Two-Stream Neural Network for AIoT-based Surveillance: Ullah et al. [75] proposed a novel two-stream neural network architecture that facilitates the real-time detection of events in AIoT-based surveillance systems. Their work emphasized the significance of this approach in resource-constrained environments, showcasing its potential for efficient event identification.

2) HUMAN–ROBOT INTERACTION (HRI)

The integration of HAR into the HRI is crucial for developing intelligent robotic systems. Mojarad et al. [76] proposed a hybrid approach that combines ontology and ML techniques,

thereby enhancing the robot's understanding and interaction capabilities within human environments. Another study by Martinez et al. [7] specifically focused on the role of HAR in robotic imitation learning, which is crucial for robots to acquire the ability to learn and mimic human activities for social interaction and assistance. Additionally, Rea et al. [77] identified key moments in human actions, which significantly contributed to more intuitive and responsive robot behavior in real-time, thereby advancing the field of HRI.

3) SILHOUETTE SEQUENCES

The integration of 3D skeletal posture estimation with 2D forms for real-time low-dimensional feature extraction and fusion is highlighted in silhouette sequences of HAR. Chaaraoui et al. [78] successfully combined skeletal and silhouette features, resulting in a significant improvement in recognition rates for capturing dynamic human actions. Elharrouss et al. [79] introduced an innovative approach that analyzed silhouettes by incorporating time as a third spatial dimension, thereby showcasing advancements in dynamic human action capture. Expanding the field, Murtaza et al. [80] and Maity et al. [81] focused on multi-view HAR and silhouette normalization to enhance accuracy and computational efficiency. Ramya and Rajeswari [82] explored the use of distance transforms and entropy for silhouette analysis and demonstrated the versatility of silhouette-based methods across various settings.

4) COMPUTATIONAL MODELING

Computational modeling in HAR centers on the interplay between the hardware capability and algorithm efficiency. Meng et al. [83] deployed Field-Programmable Gate Array (FPGA) technology to enable real-time on-device processing in HAR, which is crucial for immediate application requirements, such as surveillance. Liu et al. [84] developed bioinspired models that emulate the neural processing of the visual cortex, streamlining computational demands. Several recent studies [85], [86] presented FPGA-compatible architectures that prioritize computational frugality while maintaining accuracy, thereby broadening the scope of HAR in portable and embedded devices. Javed et al. [87] focused on the practical integration of HAR into everyday life through smartphones, thereby propelling the field towards user-centered applications.

5) GRAPH-BASED APPROACHES

Utilization HAR has proven highly valuable in various domains. Aoun et al. [88] demonstrated the effectiveness of graphs for managing spatiotemporal data and skeleton-based structures in the context of HAR. Similarly, Li and Leung [89] employed graph kernels to analyze action similarity, and achieved remarkable results in the analysis of 3D skeletal data from depth-captured benchmark datasets. Mondal et al. [90] took a step further by developing an end-to-end fast Graph Neural Network (GNN) that transforms time-series data into structural graph representations, thereby introducing a novel dimension to HAR. Ahmad et al. [91] explored various Graph Convolutional Network (GCN)-based methods, including reinforcement learning and encoder-decoder GCN models, shedding light on the continuous advancements in this field. Zhou et al. [92] applied graph-based techniques to long-term activity patterns, effectively integrating local temporal and global semantic relations to gain a comprehensive understanding of human actions.

6) DEEP LEARNING ARCHITECTURES

Various DL architectures have been used in HAR to distinguish intricate human actions. This discourse delves into multiple prominent approaches, assessing their adoption and suitability, while emphasizing their advantages and limitations.

One such approach is SlowFast, a dual-stream network capable of simultaneously capturing spatial and temporal features. This makes it well-suited for analyzing complex and diverse human motions [93]. SlowFast was further enhanced by incorporating a You Only Live Once (YOLO) model and temporal attention mechanism, enabling spatial localization and temporal alignment [94]. However, it is important to note that SlowFast has some limitations. These include high computational costs, sensitivity to hyperparameters, and challenges in effectively handling occlusions and background clutter.

I3D ResNet50 is a network that combines 3D convolution with ResNet50 feature extraction. It is designed to be adaptable to varying activity durations and has demonstrated superior performance on a kinetics dataset [95]. To make it suitable for real-time HAR on mobile devices, it was optimized by reducing its complexity using a lightweight RGB model and employing a knowledge distillation technique [96]. However, this approach has some limitations including high memory consumption, low efficiency, and poor generalization to unseen domains.

In contrast, two-stream CNN is a network that introduces separate spatial and temporal streams. This architecture has proven to be effective for recognizing diverse human actions and has achieved competitive results on the UCF-101 and HMDB-51 datasets [97]. To enhance its performance in skeleton-based HAR, it has been fused with capsule networks [98]. Despite its strengths, the two-stream CNN has some limitations. This relies heavily on optical flow computations, which are computationally expensive. In addition, its fusion efficiency is relatively low and lacks spatiotemporal coherence.

The 3D CNN with LSTM is a powerful network that combines the capabilities of 3D CNN with Long Short-Term Memory (LSTM) for spatiotemporal feature learning. This network recognizes prolonged sequential patterns and has shown impressive performance on benchmark datasets such as KTH and Weizmann [44]. It is commonly used for abnormal behavior recognition by employing a multi-scale

feature fusion module and a multitask learning scheme to enhance its capabilities [99]. However, it is important to note that this approach has certain limitations. First, it exhibits high complexity, which makes it computationally expensive and time-consuming. Second, its robustness is relatively low, making it sensitive to variations and noise in input data. Finally, it faces challenges in handling noisy and irregular actions that can affect performance in real-world scenarios.

On the other hand, transformer-based networks leverage attention mechanisms to capture long-range dependencies in temporal sequences. This network has shown great promise in modeling complex human actions and has achieved remarkable results in various natural language processing tasks [100]. Recently, efforts have been made to develop lightweight transformer models specifically designed for HAR on mobile devices [101]. These models aim to reduce the computational cost and model size while maintaining a satisfactory performance. However, it is worth mentioning that transformer-based networks also have limitations. First, they often require significant computational resources, which can be challenging for resource-constrained devices. Second, a large model can impose storage and memory constraints. Finally, similar to many DL models, transformer-based networks are vulnerable to adversarial attacks, which can compromise their performance and reliability in real-world applications.

Capsule Networks (CapsNets) are a type of neural network that focuses on capturing hierarchical representations of features. They excel in handling spatial hierarchies and have demonstrated impressive performance on popular datasets such as MNIST and CIFAR-10 [102]. CapsNets are utilized in skeleton-based HAR with Graph Convolutional Networks (GCN) [103]. However, it is important to acknowledge the limitations of CapsNets, including their high sensitivity to hyperparameters, limited scalability, and challenges in modeling temporal features.

These architectural approaches collectively contribute to the evolving landscape of HAR by offering unique strengths and by addressing specific challenges in understanding and recognizing diverse human actions. The selection of an appropriate architecture depends on the intricacies of the targeted recognition task, the characteristics of the dataset being used, and computational considerations.

A comprehensive history of research on action and behavior recognition is presented in Table 5. Although many studies have provided an overview of different methods for detecting human actions, they often overlook a detailed analysis of the advantages and disadvantages associated with each approach.

E. OPEN CHALLENGES AND LIMITATIONS

HAR systems encounter various obstacles, particularly in the realm of image analysis on a global scale, where each pixel plays a role in the final descriptor. Conventional approaches require manual detection and background subtraction, which makes them susceptible to variations in lighting, background noise, and visibility. In the realm of surveillance, video analysis is employed by law enforcement and security agencies to ensure public safety, monitor events, and resolve crime. A substantial amount of data acquired from surveillance systems necessitates algorithms that strike a balance between swift training and detection, high reliability, and the ability to learn from limited datasets [108]. The following sections discuss the common challenges faced by HAR systems.

1) DATA COLLECTION AND PRE-PROCESSING

Accurate prediction models rely heavily on effective data collection and pre-processing. In their study, Jiang et al. [109] emphasized the significance of accurately recording affected signals by considering the diversity of human activities across different locations and the influence of ambient sensors. The process of data labeling varies, as some scenarios involve manual labeling, whereas others adopt alternative approaches.

Brezmes et al. [110] used smartphones to identify and classify six distinct movements. Anguita et al. [111] employed Support Vector Machines (Support Vector Machine (SVM)) to model human behavior using data from a waist-worn smartphone inertial sensor. Kose et al. [112] tracked real-time activity using a smartphone accelerometer. Fuentes et al. [113] and Lara et al. [114] employed smartphone accelerometers for real-time motion detection. Lara et al. [114] further enhanced their processing techniques by incorporating ML methods. Lee et al. [115] addressed the issue of noise in accelerometer data by focusing on the vector magnitude of the signal.

Data collection and preprocessing pose various challenges such as the management of erroneous forms, outliers, and anomalies. Machine learning (ML) algorithms may occasionally disregard outliers, whereas the presence of duplicate frames can lead to confusion within models. It is imperative to address these concerns, as they have the potential to significantly influence decision-making processes and the overall performance of the model.

2) DATASET MODELING AND CONFIGURATION

Modeling actions from video sequences poses significant challenges. In their study, Zhang et al. [116] introduced motion context, a representation that combines image and motion information. This representation is robust to variations in action size and effectively captures the 3D characteristics of actions, leading to improved results for specific datasets. However, obstacles remain to be overcome, such as space-blind motion words (MWs) and limitations of graphical models when dealing with limited video datasets.

Advancements in multimedia and computer vision have recently focused on the detailed analysis and understanding of videos, including objects and their relationships. However, current evaluations often rely on small datasets or indirect metrics. To address this issue, Shang et al. [65] proposed a

TABLE 5. Summary of literature on HAR techniques.

Author	Contributions
[83]	SVM classifiers and using an FGPA-based video processing architecture. SVM classifiers cannot be used for large datasets, especially linear SVMs.
[104]	K-D Trees which is a special technique of K-partitioning (novelty and first 3D approach). Silhouette sequence formation is time-consuming.
[89]	Joint-Spatial Graphs (a graph with spatial features for human joints). A good approach to recognizing human actions due to quick response time but limited actions for the dataset used.
[105]	Works on training the model using the Directed Graphs approach. Great Accuracy Over A Large dataset but requires large storage space.
[74]	Trained CNNs for performance improved over Stanford 40 dataset. A single dataset with only 40 video clips seems to be a limited approach.
[69]	Recognizing actions by skeleton forming using joint movements of a human. GPU excessive workload system performance.
[85]	RGB Camera CNN NVIDIA JETSON XAVIER embedded board. & Recognizing the Front View of the person through pose estimation algorithm.
[106]	Accelerating RFC-HyPGCN on Xilinx XCKU-115 FGPA & GCN action recognition. Increased Throughput And Efficiency but dataset limited.
[107]	Optimizing the hardware architecture to shorten the latency and improve the overall performance based on Xilinx Ultrascale+ ZCU102 FGPA-based neural networks. The UCF101 dataset adds 13,320 video clips to UCF50. Optimizing the hardware architecture to shorten the latency and improve overall performance.
[93]	SlowFast is a dual-stream network that captures spatial and temporal features concurrently, suitable for complex and varied human motions. Augmented with a YOLO model and temporal attention mechanism for spatial and temporal localization. These limitations include high computational cost, sensitivity to hyperparameters, and difficulty in handling occlusions and background clutter.
[95]	I3D ResNet50 is a network that merges 3D CNN with ResNet50 feature extraction, is adaptable to various activity durations, and has superior performance on the Kinetics dataset. This is reduced by lightweight 3D CNN and knowledge distillation techniques for real-time HAR on mobile devices. These limitations include high memory consumption, low efficiency, and poor generalization to unseen domains.
[97]	Two-stream CNN is a network that introduces separate spatial and temporal streams, effective for diverse human actions, and competitive results on the UCF-101 and HMDB-51 datasets. Fused by capsule networks for skeleton-based HAR. Limitations include high dependency on optical flow computation, low fusion efficiency, and a lack of spatiotemporal coherence.
[44]	A network that combines RGB with LSTM for spatiotemporal feature learning, excels in recognizing prolonged sequential patterns, and performs well on the KTH and Weizmann datasets. Used for abnormal behavior recognition using a multiscale feature fusion module and multitask learning scheme. These limitations include high complexity, low robustness, and difficulty in handling noisy and irregular actions.
[100]	Transformer-based network uses attention mechanisms to capture long-range dependencies in temporal sequences, promising complex human actions, and remarkable results on various natural language processing tasks. Reduced by a lightweight transformer model for HAR on mobile devices. These limitations include high computational cost, large model size, and vulnerability to adversarial attacks.
[102]	CapsNets is a network that focuses on hierarchical representations of features, is proficient in handling spatial hierarchies, and has impressive results on the MNIST and CIFAR-10 datasets. used for skeleton-based HAR with a GCN. These limitations include high sensitivity to hyperparameters, low scalability, and difficulty in modeling temporal features.

cost-effective annotation pipeline. This pipeline addresses challenges related to keyframe creation, task decomposition, and other factors such as free camera motion, illumination, object deformation, and keyframe generation. By enabling large-scale annotation, this approach overcomes the limitations of previous evaluations.

3) THE ROLE OF OPEN-ACCESS AND COMMERCIAL TOOLS IN HAR

The utilization of HAR systems in both their development and deployment is significantly enhanced by the availability of open access and commercial tools. Open-access tools, such as OpenPose and DeepPose, offer cost-effective and easily accessible solutions, enabling researchers and developers to quickly prototype and experiment [117]. These tools are often equipped with pre-trained models and user-friendly interfaces, expanding the accessibility of HAR technology.

On the other hand, commercial tools such as Microsoft Kinect and Intel RealSense provide advanced capabilities and robustness for professional and large-scale applications [118]. These tools typically offer comprehensive support and integration options, making them well-suited for complex surveillance systems or commercial activity monitoring solutions.

The decision to opt for either open access or commercial tools is contingent upon various factors, including the financial resources available, scale of implementation, and specific requirements of the HAR system [119]. Open-access tools play a crucial role in facilitating research and fostering innovation. However, they may lack the ability to be tailored to specific needs and may not offer the same level of support as their commercial counterparts. However, commercial tools provide advanced functionalities but are prohibitively expensive and less adaptable for research purposes. This dichotomy poses a challenge when selecting tools for HAR scenarios because it necessitates striking a balance between advanced features, customization options, and cost-effectiveness.

4) ANALYZING A DATASET BASED ON INPUT FROM IMAGES OR VIDEO FRAMES

Video analysis in HAR involves processing a sequence of static images and video frames. The extraction and classification of frames from videos pose several challenges, including

variations between different classes, similarities within the same class, poor image quality, interactions between individuals in a group, long-distance shots, complex backgrounds, and interactions involving multiple subjects [120], [121].

Most HAR datasets consist of video clips that capture the actions of ADLs. Analyzing every frame in these videos can be computationally and memory intensive. To address this issue, Serpush and Rezaei [122] proposed a method that categorizes selected frames from a video, thereby improving processing speed and enabling real-time applications. Typically, a subset of 30 frames is considered sufficient for an accurate categorization [123]. Activities and confidence levels were determined by averaging the categorization results of the selected frames. However, this approach may encounter difficulties when dealing with videos involving multiple concurrent activities, which can complicate the recognition and categorization of actions.

5) METRICS FOR PERFORMANCE EVALUATION

More than 650 million people have disabilities worldwide. They should be provided with a monitoring system that is both reliable and precise with a high recall that helps reduce action recognition mistakes [124]. ML techniques such as Random Forest (RF), KNN, ANN, and SVM have been used for HAR [125], [126], [127]. However, an HAR system is only acceptable if it performs better in terms of the evaluation metrics. In this regard, some well-known and widely used evaluation metrics include accuracy, precision, recall, and f-score [128]. Accuracy (Equation 1) is the most common metric used to determine how well a model works by finding the correct number of predictions from the total predictions. Precision (Equation 2) is the percentage of the total number of accurate forecasts possessed by an item. Recall (Equation 3) is the percentage of real positive cases detected accurately in a given dataset. F-score/F1-score (Equation 4) evaluates binary classification systems by combining the model's precision and recalling samples as 'positive' or 'negative.'

$$Accuray = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP: True Positive, and it is expected that the observation will be positive and positive. FP: False positive: the observation is expected to be positive, but it turned out to be negative. TN: True Negatives, the prediction was that the observation would be negative, and thus it was. FN: False Negatives; instead of being negative, the observation is positive. Table 6 presents a summary of the literature on the types of datasets and the evaluation metrics used.

A sensor-based Body Area Network (BAN)-based DL) model, InnoHAR, was developed in [129] by integrating an inception neural network with a recurrent neural network. With this method, metrics such as those in the table can be evaluated using time-series features in the form of sensor waveforms from more than one channel.

TABLE 6. Summary of literature on evaluation metrics.

Authors	Dataset	Evaluation Metrics
[130]	Own data	Accuracy
[132]	Own data	Accuracy, f-measure
[129]	UCI HAR	Accuracy, f-measure
[131]	Own data	Accuracy, precision, recall
[115]	Real-World Human Action Recognition	Precision, recall, F1 Score
[132]	Own data	Accuracy, f-measure, Precision, recall

Bao and Intille [130] developed and evaluated classifiers based on accelerated data, such that the *thigh* and *wrist* were the only two pitched of the five small biaxial accelerometers available. This method was also used to measure factors such as the correlation mean and frequency-domain entropy. They found that their decision-tree-based classifier performed best in detecting ordinary activities, with an overall accuracy rate of 84 percent.

Lee et al. [115] attempted to improve accuracy and precision scores by employing an approach that lets the users' smartphones provide the data for a 1D CNN-based triaxial accelerometer model. They converted the X, Y, and Z acceleration data into vector magnitude data before using them in training a 1D CNN. Overall, 1D CNN-based ternary action recognition exceeded the baseline Random Forest approach by 89.10% for ternary action recognition.

Bayat et al. [131] proposed an HAR system using a digital low-pass mesh that isolates gravity acceleration from body acceleration and achieved 91.15% accuracy using the average of probabilities as the fusion technique.

Attal et al. [132] used chest, right thigh, and ankle sensors in combination with a preprocessing and data classification approach for supervised classification. They used SVM, KNN, RF, and mixture models [133]. They evaluated their results based on the f-measure, recall, and precision. They found that three MTx inertial IMUs were located in the chest, right thigh, and left ankle compared with 12 static, dynamic, and transition action classification methods. Unsupervised classification techniques can quickly build models from unlabeled data and are inexpensive to run on computers. However, supervised techniques are more accurate when working with raw data, or extracted or selected features. A real-world test demonstrated that RF performed better than SVM. The SLGMM is difficult to use when viewed. Additionally, the Hidden Markov Model, Gaussian Mixture Model, and K-means can handle temporal and sequential data.

Lawal and Bano [134] generated image sequences using two trained CNN sensors and combined them to forecast classes based on human activities carried out on a public dataset with an f-score of 0.87%.

TABLE 7. A summary of HAR taxonomy.

S.No.	Techniques	Application	Explanation
1.	Computational modeling	Dynamic	Real-time FPGA-based devices can recognize human actions by duplicating or adding processing cores to modify the system's processing power. Intelligent settings, human-machine communications, and security systems utilize this technology.
2.	Silhouette Sequence Point Clouds	Dynamic	This approach analyzes the time sequence of the camera silhouettes. They have built action-based spaces. The activities and shape information were recognized using 3-D point clouds. The preliminary results demonstrate that the approach can consistently detect actions, even when the data are dynamic and from numerous sources and periods.
3.	Graph-based approach	Static	Classification of human behavior based on graphs. This model maintains a complex spatial arrangement of the joints in the body by considering how they move and change over time.
4.	Human motion understanding for human–robot interaction	Dynamic	New FPGA action recognition hardware based on two-stream neural networks. This design delivers the same accuracy as existing 3CD baseline models on the Xilinx Ultrascale+ZCU102, with an order of magnitude fewer operations.
5.	Graph-based approach	Static	This study categorizes numerous GCN models to address the graph representation learning challenge in computer vision-related applications.
6.	Computational modeling	Dynamic	This technique looks for "dynamic instants" in an action's timeline to determine when a partner's movements begin, end, or change.
7.	Neural Networks	Dynamic	Identify typical 3D actions in NTU RGB D by building a movement-based interactive system using tree topologies. They found that bones and joints represent the tree's base, whereas child nodes that link provide incoming and outgoing edges.
8.	Pre-trained CNNs	Dynamic	Combines class-based and instance-based success rates to assess traditional and ABC-optimized transfer models (test data success rate). All class- and instance-based NASNet-Large parameterize the ABC-optimized CNN.

IV. CONCLUSION AND FUTURE WORK

The impact of HAR on daily life, real-time situations, and collaborative efforts has been discussed extensively in the literature. This paper offers a comprehensive review of the current state of HAR systems. The findings from the various sources cited in this paper highlight several key points. In addition, a summary of the taxonomy of HAR techniques and methods is presented in Table 7.

Researchers have focused on various aspects of HAR, including user behavior, static and dynamic living activities, and lifestyles. However, limited attention has been paid to real-time HAR for medical security. The complexity of real-time actions, hardware and technical limitations, and data scarcity pose challenges for the development of HAR systems. Meeting the demands of real-time public or occupied spaces and adapting to changing scenery requires significant computing power. Furthermore, the availability of high-quality real-time data is limited, and existing literature falls short of addressing real-time activities. For instance, Kinect devices, with their limited field of view, can only detect falls and unusual behavior in smart homes.

Privacy concerns arise in vision-based HAR systems that rely on cameras. Some individuals may be reluctant to have their data, including images and videos, stored permanently, raising privacy concerns.

Sensor-based HAR has advantages and disadvantages similar to those of other technologies. Some sensors are worn by individuals, whereas others are embedded in buildings and vehicles. Wearing body-implanted sensors may cause discomfort; however, they provide versatile and detailed motion data. In contrast, wearable sensors offer the advantage of capturing real-time data despite potential HAR data loss.

The widespread use of smartphones has made HAR sensors readily available for the general population. This accessibility opens opportunities for novel combinations of devices and sensors in HAR research.

In the literature, several HAR implementations utilize both ML and DL methods. Researchers have suggested the use of hybrid HAR approaches. The following are several potential areas that can be explored in future research.

- Researchers involved in HAR should focus on addressing two key aspects: (1) acquiring sufficient data for HAR systems and (2) evaluating their performance effectively.
- Previous evaluations have primarily concentrated on specific aspects of HAR, such as an integrated HAR with artificial intelligence frameworks [135] or the categorization of video frames [122] to enhance real-time processing. To provide valuable and accurate insights into human behavior and interactions with the

environment, it is recommended to establish connections between HAR and Computer Vision.

- While the KTH and Weizmann datasets have been extensively studied and utilized due to their simplicity and limited number of videos, the MSR dataset, which captures behavior from multiple cameras, has been identified as a valuable benchmark dataset. However, further research is necessary to develop more suitable datasets that encompass a wider range of activities and perspectives, thereby enabling the training of state-of-the-art DL models because they span a broader range of activities and perspectives.

This review article aims to establish a foundation for further investigation into significant issues and potential advancements in the field.

LIST OF ABBREVIATIONS

3D CNN	Three-Dimensional Convolutional Neural Network
ADL	Activities of Daily Living
BAN	Body Area Network
CNN	Convolutional Neural Network
DKN	Dynamic Kernel Network
DL	Deep Learning
FPGA	Field-Programmable Gate Array
GCN	Graph Convolutional Network
HAR	Human Action Recognition
HCI	Human-Computer Interaction
ML	Machine Learning
RF	Random Forest
RGB	Red Green Blue
SDF	Signed Distance Function
SVLRM	Subspace Video Linear Regression Model
SVM	Support Vector Machine
TOF	Time-of-Flight
YOLO	You Only Live Once

REFERENCES

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] E. J. Amirbandi and G. Shamsipour, "Exploring methods and systems for vision based human activity recognition," in *Proc. 1st Conf. Swarm Intell. Evol. Comput. (CSIEC)*, Mar. 2016, pp. 160–164.

[3] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Müller, "Deep learning for time series classification: A review," *Data Min. Knowl. Disc.*, vol. 33, pp. 917–963, Mar. 2019.

[4] M. Agarwal, L. Saba, S. K. Gupta, A. Carriero, Z. Falaschi, A. Paschè, P. Danna, A. El-Baz, S. Naidu, and J. S. Suri, "A novel block imaging technique using nine artificial intelligence models for COVID-19 disease classification, characterization and severity measurement in lung computed tomography scans on an Italian cohort," *J. Med. Syst.*, vol. 45, no. 3, pp. 1–30, Mar. 2021.

[5] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[6] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.

[7] L. Martînez-Villaseñor and H. Ponce, "A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 6, 2019, Art. no. 1550147719853987.

[8] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015.

[9] H. Paulheim, "Generating possible interpretations for statistics from linked open data," in *Proc. Extended Semantic Web Conf.* Cham, Switzerland: Springer, 2012, pp. 560–574.

[10] J.-L. Chung, L.-Y. Ong, and M.-C. Leow, "Comparative analysis of skeleton-based human pose estimation," *Future Internet*, vol. 14, no. 12, p. 380, Dec. 2022.

[11] G. Diraco, G. Rescio, A. Caroppo, A. Manni, and A. Leone, "Human action recognition in smart living services and applications: Context awareness, data availability, personalization, and privacy," *Sensors*, vol. 23, no. 13, p. 6040, Jun. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/6040

[12] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: A survey," *Neural Comput. Appl.*, vol. 35, no. 5, pp. 4145–4182, Feb. 2023.

[13] R. S. Ransing and M. Rajput, "Smart home for elderly care, based on wireless sensor network," in *Proc. Int. Conf. Nascent Technol. Eng. Field (ICNTE)*, Jan. 2015, pp. 1–5.

[14] B. Jo and S. Kim, "Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices," *Traitement Signal*, vol. 39, no. 1, pp. 119–124, Feb. 2022.

[15] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, S. A. Velastin, and P. Zegers, "A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera," *Sensors*, vol. 20, no. 7, p. 1825, Mar. 2020.

[16] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. Tianyi Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, pp. 287–304, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025518309964

[17] J. Gong, R. Li, H. Yao, X. Kang, and S. Li, "Recognizing human daily activity using social media sensors and deep learning," *Int. J. Environ. Res. Public Health*, vol. 16, no. 20, p. 3955, Oct. 2019.

[18] R. G. Guendel, "Further investigation of two-way classification for activities of daily living," in *Proc. 17th Eur. Radar Conf. (EuRAD)*, Jan. 2021, pp. 210–213.

[19] R. Saini, P. Kumar, P. P. Roy, and D. P. Dogra, "A novel framework of continuous human-activity recognition using Kinect," *Neurocomputing*, vol. 311, pp. 99–111, Oct. 2018.

[20] A. G. Salguero and M. Espinilla, "Ontology-based feature generation to improve accuracy of activity recognition in smart environments," *Comput. Electr. Eng.*, vol. 68, pp. 1–13, May 2018.

[21] R. Chen, Y. Rao, R. Cai, X. Shi, Y. Wang, and Y. Zou, "Design and implementation of human-computer interaction based on user experience for dynamic mathematics software," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 428–433.

[22] A. Stenila and M. M. Asuntha, "Human computer interaction based HEMD using hand gesture," *Int. J. Adv. Eng., Manage. Sci.*, vol. 3, no. 5, pp. 587–590, 2017.

[23] L. Maradani, "Human activity recognition," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 7, pp. 1983–1988, Jul. 2022.

[24] R. Heimgärtner, *Culturally-Aware HCI Systems*. Cham, Switzerland: Springer, 2018, pp. 11–37.

[25] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, and C. Sminchisescu, "BlazePose GHUM holistic: Real-time 3D human landmarks and pose estimation," 2022, *arXiv:2206.11678*.

[26] A. Rebelo, D. V. Martinho, J. Valente-Dos-Santos, M. J. Coelho-E-Silva, and D. S. Teixeira, "From data to action: A scoping review of wearable technologies and biomechanical assessments informing injury prevention strategies in sport," *BMC Sports Sci., Med. Rehabil.*, vol. 15, no. 1, pp. 1–14, Dec. 2023.

[27] P. S. Glazier and S. Mehdizadeh, "Challenging conventional paradigms in applied sports biomechanics research," *Sports Med.*, vol. 49, no. 2, pp. 171–176, Feb. 2019.

[28] K. Singh, M. Rastogi, M. Mahajan, and D. R. Kumar, "Accident prevention by detecting drivers fatigueness," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 5, pp. 3919–3924, May 2022.

[29] J.-J. Wan, Z. Qin, P.-Y. Wang, Y. Sun, and X. Liu, "Muscle fatigue: General understanding and treatment," *Experim. Mol. Med.*, vol. 49, no. 10, p. 384, Oct. 2017.

[30] D. N. Jha, Z. Chen, S. Liu, M. Wu, J. Zhang, G. Morgan, R. Ranjan, and X. Li, "A hybrid accuracy- and energy-aware human activity recognition model in IoT environment," *IEEE Trans. Sustain. Comput.*, vol. 8, no. 1, pp. 1–14, Jan. 2023.

[31] P. Kumar and S. Suresh, "How tri-axial sensors influenced the location-based heterogeneous activities recognition rates: An exploratory analysis," in *Proc. 2nd Int. Conf. Range Technol. (ICORT)*, Aug. 2021, pp. 1–6.

[32] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021.

[33] J. Alderden, K. P. Drake, A. Wilson, J. Dimas, M. R. Cummins, and T. L. Yap, "Hospital acquired pressure injury prediction in surgical critical care patients," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, p. 12, Dec. 2021.

[34] H. Jo, W. Lee, and E. Kim, "Mixture density-PoseNet and its application to monocular camera-based global localization," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 388–397, Jan. 2021.

[35] G. Ding, Q. Wu, L. Zhang, Y. Lin, T. A. Tsiftsis, and Y.-D. Yao, "An amateur drone surveillance system based on the cognitive Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 29–35, Jan. 2018.

[36] P. K. Roy and H. Om, "Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos," in *Advances in Soft Computing and Machine Learning in Image Processing*. Cham, Switzerland: Springer, 2018, pp. 277–294.

[37] C.-B. Jin, S. Li, T. D. Do, and H. Kim, "Real-time human action recognition using CNN over temporal images for static video surveillance cameras," in *Proc. Pacific Rim Conf. Multimedia*. Germany: Springer-Verlag, 2015, pp. 330–339.

[38] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 838–845.

[39] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jul. 2004, pp. 819–826.

[40] R. Kumar and S. Kumar, "Survey on artificial intelligence-based human action recognition in video sequences," *Opt. Eng.*, vol. 62, no. 2, Feb. 2023, Art. no. 023102.

[41] U. M. Kamthe and C. G. Patil, "Suspicious activity recognition in video surveillance system," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–6.

[42] P. Khaire and P. Kumar, "Deep learning and RGB-D based human action, human–human and human-object interaction recognition: A survey," *J. Vis. Commun. Image Represent.*, vol. 86, Jul. 2022, Art. no. 103531.

[43] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 3, Aug. 2004, pp. 32–36.

[44] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Human Behavior Understanding*. Berlin, Germany: Springer, 7065, pp. 29–39.

[45] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.

[46] L. Cai, C. Liu, R. Yuan, and H. Ding, "Human action recognition using lie group features and convolutional neural networks," *Nonlinear Dyn.*, vol. 99, no. 4, pp. 3253–3263, Mar. 2020.

[47] A. Ben Tamou, L. Ballihi, and D. Aboutajdine, "Automatic learning of articulated skeletons based on mean of 3D joints for efficient action recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 04, Apr. 2017, Art. no. 1750008.

[48] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 479–485.

[49] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, "Spatial-temporal data augmentation based on LSTM autoencoder network for skeleton-based human action recognition," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3478–3482.

[50] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing RGB and depth map features for human activity recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2012, pp. 1–4.

[51] Y. He, S. Shirakabe, Y. Satoh, and H. Kataoka, "Human action recognition without human," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 11–17.

[52] S. M. Vaniya and B. Bharathi, "Exploring object segmentation methods in visual surveillance for human activity recognition," in *Proc. Int. Conf. Global Trends Signal Process., Inf. Comput. Commun. (ICGTSPICC)*, Dec. 2016, pp. 520–525.

[53] A. N. Shuaibu, A. S. Malik, I. Faye, and Y. S. Ali, "Pedestrian group attributes detection in crowded scenes," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, pp. 1–5.

[54] S. Deep and X. Zheng, "Leveraging CNN and transfer learning for vision-based human activity recognition," in *Proc. 29th Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, Nov. 2019, pp. 1–4.

[55] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: A survey," *Multimedia Tools Appl.*, vol. 79, nos. 41–42, pp. 30509–30555, Nov. 2020.

[56] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "AVID: Adversarial visual irregularity detection," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 488–505.

[57] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.

[58] S. Antoshchuk, M. Kovalenko, and J. Sieck, "Gesture recognition-based human–computer interaction interface for multimedia applications," in *Digitisation of Culture: Namibian and International Perspectives*. Cham, Switzerland: Springer, 2018, pp. 269–286.

[59] N. Golestani and M. Moghaddam, "Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks," *Nature Commun.*, vol. 11, no. 1, pp. 1–11, Mar. 2020.

[60] A. Haria, A. Subramanian, N. Asokkumar, S. Poddar, and J. S. Nayak, "Hand gesture recognition for human computer interaction," *Proc. Comput. Sci.*, vol. 115, pp. 367–374, Jan. 2017.

[61] A. Poulos, C. Brown, D. McCulloch, and J. Cole, "Context-aware augmented reality object commands," U.S. Patent 10 705 602, Jul. 7, 2020.

[62] K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: A survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, Jun. 2017.

[63] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "ViHASi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods," in *Proc. 2nd ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Sep. 2008, pp. 1–10.

[64] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, p. 1101, Oct. 2017.

[65] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 279–287.

[66] K. H. Walse, R. V. Dharaskar, and V. M. Thakare, "A study of human activity recognition using AdaBoost classifiers on WISDM dataset," *Inst. Integrative Omics Appl. Biotechnol. J.*, vol. 7, no. 2, pp. 68–76, 2016.

[67] A. Jain and V. Kanhangad, "Gender classification in smartphones using gait information," *Expert Syst. Appl.*, vol. 93, pp. 257–266, Mar. 2018.

[68] L. Wang, W. Li, W. Li, and L. van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.

[69] M. S. Khan, A. Ware, M. Karim, N. Bahoo, and M. J. Khalid, "Skeleton based human action recognition using a structured-tree neural network," *Eur. J. Eng. Res. Sci.*, vol. 5, no. 8, pp. 849–854, Aug. 2020.

[70] D. Burns, P. Boyer, C. Arrowsmith, and C. Whyne, "Personalized activity recognition with deep triplet embeddings," *Sensors*, vol. 22, no. 14, p. 5222, Jul. 2022.
[71] L. Xu, W. Yang, Y. Cao, and Q. Li, "Human activity recognition based on random forests," in *Proc. 13th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Jul. 2017, pp. 548–553.
[72] S. Gupta, "Deep learning based human activity recognition (HAR) using wearable sensor data," *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 2, Nov. 2021, Art. no. 100046.
[73] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A novel approach for robust multi human action recognition and summarization based on 3D convolutional neural networks," 2019, *arXiv:1907.11272*.
[74] T. Ozcan and A. Basturk, "Performance improvement of pre-trained convolutional neural networks for action recognition," *Comput. J.*, vol. 64, no. 11, pp. 1715–1730, Nov. 2019.
[75] W. Ullah, A. Ullah, T. Hussain, K. Muhammad, A. A. Heidari, J. del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data," *Future Gener. Comput. Syst.*, vol. 129, pp. 286–297, Apr. 2022.
[76] R. Mojarad, F. Attal, A. Chibani, S. R. Fiorini, and Y. Amirat, "Hybrid approach for human activity recognition by ubiquitous robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 5660–5665.
[77] F. Rea, A. Vignolo, A. Sciutti, and N. Noceti, "Human motion understanding for selecting action timing in collaborative human–robot interaction," *Frontiers Robot. AI*, vol. 6, p. 58, Jul. 2019.
[78] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta, "Fusion of skeletal and Silhouette-based features for human action recognition with RGB-D devices," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 91–97.
[79] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Appl. Intell.*, vol. 51, no. 2, pp. 690–712, Feb. 2021.
[80] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Comput. Vis.*, vol. 10, no. 7, pp. 758–767, Oct. 2016.
[81] S. Maity, A. Chakrabarti, and D. Bhattacharjee, "Robust human action recognition using AREI features and trajectory analysis from silhouette image sequence," *IETE J. Res.*, vol. 65, no. 2, pp. 236–249, Mar. 2019.
[82] P. Ramya and R. Rajeswari, "Human action recognition using distance transform and entropy based features," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 8147–8173, Mar. 2021.
[83] H. Meng, M. Freeman, N. Pears, and C. Bailey, "Real-time human action recognition on an embedded, reconfigurable video processing architecture," *J. Real-Time Image Process.*, vol. 3, no. 3, pp. 163–176, Sep. 2008.
[84] H. Liu, N. Shu, Q. Tang, and W. Zhang, "Computational model based on neural network of visual cortex for human action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1427–1440, May 2018.
[85] J. Lee and B. Ahn, "Real-time human action recognition with a low-cost RGB camera and mobile robot platform," *Sensors*, vol. 20, no. 10, p. 2886, May 2020.
[86] H. Fan, C. Luo, C. Zeng, M. Ferianc, Z. Que, S. Liu, X. Niu, and W. Luk, "F-E3D: FPGA-based acceleration of an efficient 3D convolutional neural network for human action recognition," in *Proc. IEEE 30th Int. Conf. Application-Specific Syst., Architectures Processors (ASAP)*, Jul. 2019, pp. 1–8.
[87] A. R. Javed, M. U. Sarwar, S. Khan, C. Iwendi, M. Mittal, and N. Kumar, "Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition," *Sensors*, vol. 20, no. 8, p. 2216, Apr. 2020.
[88] N. Ben Aoun, M. Mejdoub, and C. Ben Amar, "Graph-based approach for human action recognition using spatio-temporal features," *J. Vis. Commun. Image Represent.*, vol. 25, no. 2, pp. 329–338, Feb. 2014.
[89] M. Li and H. Leung, "Graph-based approach for 3D human skeletal action recognition," *Pattern Recognit. Lett.*, vol. 87, pp. 195–202, Feb. 2017.
[90] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja, and R. Sarkar, "A new framework for smartphone sensor-based human activity recognition using graph neural network," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11461–11468, May 2021.
[91] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, Apr. 2021.
[92] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng, "Graph-based high-order relation modeling for long-term action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8980–8989.
[93] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[94] J.-O. Jeong, W. Hu, and P. Lalanda, "Human activity recognition with computer vision," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep., 2020. [Online]. Available: http://cs230.stanford.edu/projects_fall_2020/reports/55772236.pdf
[95] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
[96] Y. Zhao, W. Hu, and X. Hu, "Real-time human activity recognition using ResNet and 3D convolutional neural networks," in *Proc. 40th Chin. Control Conf. (CCC)*, Sep. 2021, pp. 5845–5855.
[97] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.
[98] L. Llopis-Ibor, A. Cuesta-Infante, C. Beltran-Royo, and J. J. Pantrigo, "Human activity recognition with capsule networks," in *Advances in Artificial Intelligence*. Cham, Switzerland: Springer, 2021, pp. 75–85.
[99] Y. Guan, W. Hu, and X. Hu, "Abnormal behavior recognition using 3D-CNN combined with LSTM," *Multimedia Tools Appl.*, vol. 80, no. 13, pp. 18787–18801, May 2021.
[100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.
[101] K. Sannara, F. Portet, and P. Lalanda, "Lightweight transformers for human activity recognition on mobile devices," 2022, *arXiv:2209.11750*.
[102] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2017, pp. 3856–3866.
[103] H. Damirchi, R. Khorrambakht, and H. D. Taghirad, "ARC-Net: Activity recognition through capsules," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Miami, FL, USA, 2020, pp. 1382–1388, doi: 10.1109/ICMLA51294.2020.00215.
[104] R. B. Rusu, J. Bandouch, Z. C. Marton, N. Blodow, and M. Beetz, "Action recognition in intelligent environments using point cloud features extracted from silhouette sequences," in *Proc. 17th IEEE Int. Symp. Robot Human Interact. Commun.*, Aug. 2008, pp. 267–272.
[105] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7904–7913.
[106] D. Wen, J. Jiang, J. Xu, K. Wang, T. Xiao, Y. Zhao, and Y. Dou, "RFC-HyPGCN: A runtime sparse feature compress accelerator for skeleton-based GCNs action recognition model with hybrid pruning," in *Proc. IEEE 32nd Int. Conf. Application-Specific Syst., Architectures Processors (ASAP)*, Jul. 2021, pp. 33–40.
[107] J.-M. Lin, K.-T. Lai, B.-R. Wu, and M.-S. Chen, "Efficient two-stream action recognition on FPGA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3070–3074.
[108] A. Wilkowski, M. Stefańczyk, and W. Kasprzak, "Training data extraction and object detection in surveillance scenario," *Sensors*, vol. 20, no. 9, p. 2689, May 2020.
[109] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.* New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 289–304, doi: 10.1145/3241539.3241548.
[110] T. Brezmes, J.-L. Gorricho, and J. Cotrina, "Activity recognition from accelerometer data on a mobile phone," in *Proc. Int. Work-Conf. Artif. neural Netw.* Cham, Switzerland: Springer, 2009, pp. 796–799.

[111] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Int. Workshop Ambient Assist. Living*. Cham, Switzerland: Springer, 2012, pp. 216–223.
[112] M. Kose, O. D. Incel, and C. Ersoy, "Online human activity recognition on smart phones," in *Proc. Workshop Mobile Sens., Smartphones Wearables Big Data*, vol. 16, 2012, pp. 11–15.
[113] D. Fuentes, L. Gonzalez-Abril, C. Angulo, and J. A. Ortega, "Online motion recognition using an accelerometer in a mobile device," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2461–2465, Feb. 2012.
[114] Ó. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervas. Mobile Comput.*, vol. 8, no. 5, pp. 717–729, Oct. 2012.
[115] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 131–134.
[116] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2008, pp. 817–829.
[117] S. Majumder and N. Kehtarnavaz, "Vision and inertial sensing fusion for human action recognition: A review," *IEEE Sensors J.*, vol. 21, no. 3, pp. 2454–2467, Feb. 2021.
[118] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, Feb. 2023.
[119] H. Munir, P. Runeson, and K. Wnuk, "Open tools for software engineering: Validation of a theory of openness in the automotive industry," in *Proc. Eval. Assessment Softw. Eng.*, 2019, pp. 2–11.
[120] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges and future prospects," 2022, *arXiv:2202.03274*.
[121] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.
[122] F. Serpush and M. Rezaei, "Complex human action recognition using a hierarchical feature reduction and deep learning-based method," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–15, Apr. 2021.
[123] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimedia Inf. Retr.*, vol. 2, no. 2, pp. 73–101, Jun. 2013.
[124] A. Hayat, F. Morgado-Dias, B. Bhuyan, and R. Tomar, "Human activity recognition for elderly people using machine and deep learning approaches," *Information*, vol. 13, no. 6, p. 275, May 2022.
[125] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1–21, May 2021.
[126] I. A. Bustoni, I. Hidayatulloh, A. M. Ningtyas, A. Purwaningsih, and S. N. Azhari, "Classification methods performance on human activity recognition," *J. Phys., Conf. Ser.*, vol. 1456, no. 1, Jan. 2020, Art. no. 012027.
[127] A. Laios, A. Gryparis, D. DeJong, R. Hutson, G. Theophilou, and C. Leach, "Predicting complete cytoreduction for advanced ovarian cancer patients using nearest-neighbor models," *J. Ovarian Res.*, vol. 13, no. 1, pp. 1–8, Dec. 2020.
[128] N. Amir, F. Jabeen, Z. Ali, I. Ullah, A. U. Jan, and P. Kefalas, "On the current state of deep learning for news recommendation," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1101–1144, Feb. 2023.
[129] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.
[130] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. Int. Conf. Pervasive Comput.* Berlin, Germany: Springer, 2004, pp. 1–17.
[131] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Proc. Comput. Sci.*, vol. 34, pp. 450–457, Jan. 2014.
[132] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, Dec. 2015.
[133] L. J. Chmielewski, R. Kozera, A. Orłowski, K. Wojciechowski, A. M. Bruckstein, and N. Petkov, *Computer Vision and Graphics: International Conference*, vol. 11114. Cham, Switzerland: Springer, 2018.
[134] I. A. Lawal and S. Bano, "Deep human activity recognition using wearable sensors," in *Proc. 12th ACM Int. Conf. Pervasive Technol. Rel. Assistive Environments*, Jun. 2019, pp. 45–48.
[135] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: A narrative review," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4755–4808, Aug. 2022.

MISHA KARIM received the bachelor's degree in computer science from UET Taxila. She is currently pursuing the M.S. degree in information technology with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST). During the bachelor's degree, she conducted research on HAR using 3D point clouds and ML-based convolutional networks. She accomplished a range of projects, among them the cargo management systems, the open innovation platform, the data monetization systems, the blockchain-based solution, and the metaverse. She has experience as a technical writer and a research assistant for delivering IT, business, and education projects. She has published articles on skeleton-based human-action recognition. Her research interests include HAR, web development, and DL. For more information, visit her LinkedIn profile: www.linkedin.com/in/misha-karim.

SHAH KHALID received the M.S. degree from the University of Peshawar, Pakistan, and the Ph.D. degree from Jiangsu University, China. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science (SEECS), National University of Science and Technology (NUST), Islamabad, Pakistan. He has been involved in numerous research projects in Pakistan and abroad. His research interests include information retrieval, web search engines, scholarly retrieval systems, recommender systems, knowledge graphs, social web, real-time sentiment analysis, web engineering, text summarization, federated search, and digital libraries. For more information, please visit his website at https://sites.google.com/view/shahkhalid.

ALIYA ALERYANI received the B.S. degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, the M.S. degree in computer science from Middle Tennessee State University, Murfreesboro, TN, USA, and the Ph.D. degree from the University of East Anglia, Norwich, U.K. She is currently an Assistant Professor with the College of Computer Science, King Khalid University. Her research interests include handling uncertainty in artificial intelligence and deploying artificial intelligence in sustainable development.

JAWAD KHAN received the master's degree in computer science from Kohat University of Science and Technology (KUST), Pakistan, and the Ph.D. degree in computer engineering from Kyung Hee University (Global Campus), South Korea. He is an Assistant Professor with the Department of AI Software, Gachon University (Global Campus), South Korea. He worked three years as a Post-Doctoral Researcher at the Department of Robotics, Hanyang University (ERICA Campus), South Korea. His research interests include natural language processing, information retrieval, sentiment analysis/opinion mining, text processing, social media mining, artificial intelligence, machine learning, deep learning, and computer vision.

IRFAN ULLAH received the B.S. degree (Hons.) in computer science from the Department of Computer Science, University of Malakand, Pakistan, and the M.S. and Ph.D. degrees in computer science specializing in the area of web engineering from the Department of Computer Science, University of Peshawar, Pakistan. He is currently an Assistant Professor and the Head of the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Pakistan. He has more than 14 years of teaching and research experience. He is the author of 45 research papers published in national and international journals and conferences. His research interests include information retrieval, interactive information retrieval, information service engineering, artificial intelligence, DL, recommender systems, web semantics, linked open data, ontology engineering, social web, and social book search.

ZAFAR ALI received the M.Sc. degree in computer science and the M.S. degree in web engineering from the University of Peshawar, in 2011 and 2017, respectively, and the Ph.D. degree in computer science and engineering from Southeast University, China. He is currently a Postdoctoral Fellow with the School of Computer Science and Engineering, Southeast University. He has published more than 13 research papers in reputed conferences and SCI journals. His research interests include recommender systems, information retrieval, natural language processing, graph embedding, and DL. He is a reviewer of different prestigious journals and conferences, including *Knowledge-Based Systems*, *AI Review*, *Information Fusion*, *Scientometrics*, *Soft Computing*, *Information Processing and Management*, and CIKM.