Computers in Biology and Medicine 149 (2022) 106060

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects

Md. Milon Islam [a,*], Sheikh Nooruddin [a], Fakhri Karray [a,b], Ghulam Muhammad [c]

[a] Centre for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, ON, N2L 3G1, Canada
[b] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[c] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Human activity recognition
Convolutional neural network
Multimodal sensing devices
Smartphone data
Radar signal
Vision systems

ABSTRACT

Human Activity Recognition (HAR) plays a significant role in the everyday life of people because of its ability to learn extensive high-level information about human activity from wearable or stationary devices. A substantial amount of research has been conducted on HAR and numerous approaches based on deep learning have been exploited by the research community to classify human activities. The main goal of this review is to summarize recent works based on a wide range of deep neural networks architecture, namely convolutional neural networks (CNNs) for human activity recognition. The reviewed systems are clustered into four categories depending on the use of input devices like multimodal sensing devices, smartphones, radar, and vision devices. This review describes the performances, strengths, weaknesses, and the used hyperparameters of CNN architectures for each reviewed system with an overview of available public data sources. In addition, a discussion of the current challenges to CNN-based HAR systems is presented. Finally, this review is concluded with some potential future directions that would be of great assistance for the researchers who would like to contribute to this field. We conclude that CNN-based approaches are suitable for effective and accurate human activity recognition system applications despite challenges including availability of data regarding composite or group activities, high computational resource requirements, data privacy concerns, and edge computing limitations. For widespread adaptation, future research should be focused on more efficient edge computing techniques, datasets incorporating contextual information with activities, more explainable methodologies, and more robust systems.

1. Introduction

The purpose of human activity recognition is to recognize the physical tasks of a particular individual or a group of individuals depending on the nature of the application. A few of these tasks may be carried out by a particular individual like running, jumping, walking, and sitting through changes in the entire body [1,2]. Through a particular body part movement, some activities are performed like making hand gestures [3,4]. Some cases may be done by communicating with objects, such as cooking food in the kitchen [5,6]. Any abnormal activities like sudden falls [7] are also referred to as HAR. The major successful applications of HAR include ambient assistive living [8,9], nursing home [10], health monitoring [11], rehabilitation activities [12], surveillance [13], and human-computer interaction [14]. Due to the diverse applications of HAR, it has become a very prominent research issue in the research community. Generally, HAR falls into two categories depending on the types of data such as vision-based HAR [15] and sensor-based HAR [16]. A vision-based technique analyzes the camera data as a video [17], or image [18] format while a sensor-based system interprets the sensors (accelerometer, gyroscope, radar, and magnetometer) data as a time series [19] form. Among the available sensors, the accelerometer is mostly used for HAR due to its low cost, small size, and portability [20,21]. Object sensor like Radio frequency identifier (RFID) tags [22,23] are utilized in the home environment although it is difficult to deploy. The researches have shown that sensor-based HAR [24] is quite convenient and maintains more privacy compared to vision-based HAR [25]. In addition, vision-based HAR [26,27] is more influenced by environmental factors like camera angle, lighting, and overlap between individuals, although it is less expensive to develop.

* Corresponding author.
E-mail addresses: milonislam@uwaterloo.ca (Md.M. Islam), sheikh.nooruddin@uwaterloo.ca (S. Nooruddin), karray@uwaterloo.ca (F. Karray), ghulam@ksu.edu.sa (G. Muhammad).

https://doi.org/10.1016/j.compbiomed.2022.106060
Received 6 June 2022; Received in revised form 9 August 2022; Accepted 27 August 2022
Available online 1 September 2022

Recently, deep learning (DL) algorithms have become more popular due to their automatic feature extraction capability from vision or image data [28] as well as time-series data [29] that enables the learning of high-level and meaningful features. Deep learning techniques have been generally outperforming traditional machine learning (ML) approaches for activity recognition in terms of classification performance measures such as accuracy, precision, recall, and F1 Score [30,31]. The recognition of human behavior based on deep learning architecture, especially CNN is a composite system that is comprised of several key stages. An overall system architecture of a typical CNN-based activity recognition system is illustrated in Fig. 1. The first stage is comprised of the selection and implementation of sensing devices. Data collection is the next step where an edge device is used to perceive data from input devices and transfer it to the main server through various communication systems such as Wi-Fi and Bluetooth. The deployment of computing and storage resources at the point where data is being collected and processed is referred to as edge computing which incorporates sensors for data perception as well as edge servers for reliable real-time information processing [32,33]. The feature extraction and selection stage extract the necessary features from the raw signals; this stage is performed automatically in the case of CNN; no hand-crafted feature extractions are required. This stage contains the CNN architecture or variants of CNN structure for the recognition of activities. The last step includes a notification system through which an agent (human or machine) can be notified. The notification system can aid in emergency scenarios by notifying emergency contact personnel or emergency services.

Several surveys have been conducted in this literature to highlight the recent progress in the research area of human activity recognition. The existing surveys focused on different approaches including the areas of deep learning [34–36], machine learning [37–39], sensor [40–42], and vision [26,43,44] for human activity recognition. Wang et al. [34] reviewed the existing literature for HAR considering the deep learning models, sensor modalities, and application perspectives. Although this review demonstrated different architectures and datasets, it lacks some important aspects of HAR including technical details of the type of activities, and performance measures for deep learning architectures. Additionally, the details description of the location and orientation of sensors in the datasets is not properly analyzed in this review. Ramanujam et al. [35] categorized the deep learning models into three categories: CNN, LSTM, and hybrid techniques to describe the progress

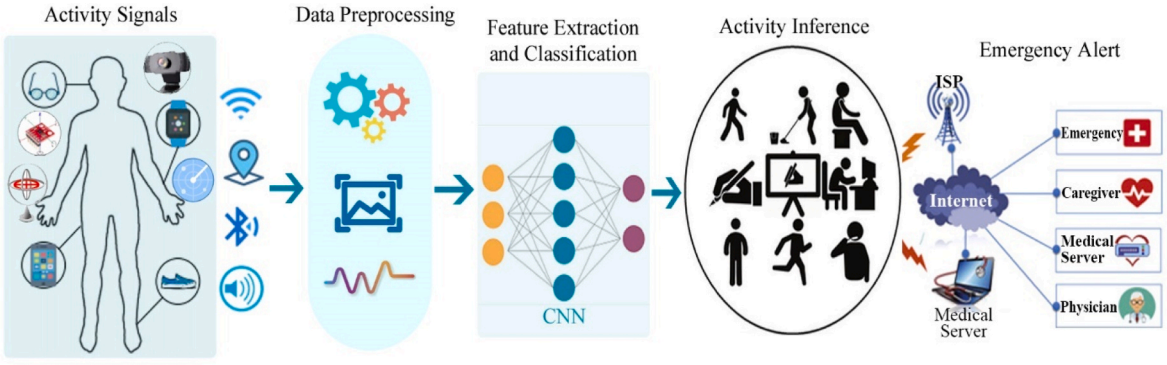

Fig. 1. An overall system architecture of CNN-based human activity recognition.

Table 1
Comparative analysis of existing state-of-the-art for human activity recognition.

Authors	Year	Major focus	Data modalities	Used keywords	Number of articles reviewed	Period
Dhillon et al. [36]	2017	Deep learning	Vision	DL, activity recognition, video, motion	24	2010–2017
Kumari et al. [40]	2017	Wearable technology	Wearable sensors	Biomedical, human computer interface, human activity monitoring, multimodal interface, shared control architecture, smart sensors, wearable sensors.	85	2005–2016
Cornacchia et al. [42]	2017	Sensor technology	Wearable sensors	Wearable, sensors, survey, activity detection, activity classification, monitoring	92	2001–2016
Bux et al. [44]	2017	Computer vision	Vision	Computer vision, HAR, objects segmentation, feature extraction, action recognition, review	76	2003–2015
Ramamurthy and Roy [37]	2018	Machine learning, data mining	Sensor	DL, ML, active learning, data mining, transfer learning, activity recognition, wearable sensors	87	2007–2018
Wang et al. [34]	2019	Deep learning	Sensor	DL, activity recognition, pattern recognition, pervasive computing	77	2013–2019
Verma et al. [38]	2019	Supervised and unsupervised machine learning	Vision	Abnormal activity, intelligent surveillance, Object detection, Object tracking, supervised and unsupervised learning	22	2005–2017
Alrazzak and Alhalabi [41]	2019	Wearable technology	Wearable sensors	–	17	2008–2018
Zhang et al. [26]	2019	Computer vision	Vision	Action detection, action feature, HAR, human-object interaction recognition, systematic survey	52	2012–2019
Beddiar et al. [43]	2020	Computer vision	Vision	HAR, behavior understanding, action representation, action detection, computer Vision, survey	40	2010–2019
Biswal et al. [39]	2021	Machine learning	Wearable sensors, vision	ML classifiers, HAR, sensor-based, vision-based	14	2008–2020
Ramanujam et al. [35]	2021	Deep learning	Wearable sensors, smartphone	DL, ML, activity recognition, wearable sensors, smartphones, context-aware	38	2015–2021
Ours	–	Convolutional neural networks	Multimodal sensing devices, smartphone, radar, and vision data	HAR, CNN, multimodal sensing devices, smartphone data, radar signal, vision systems	63	2013–2022

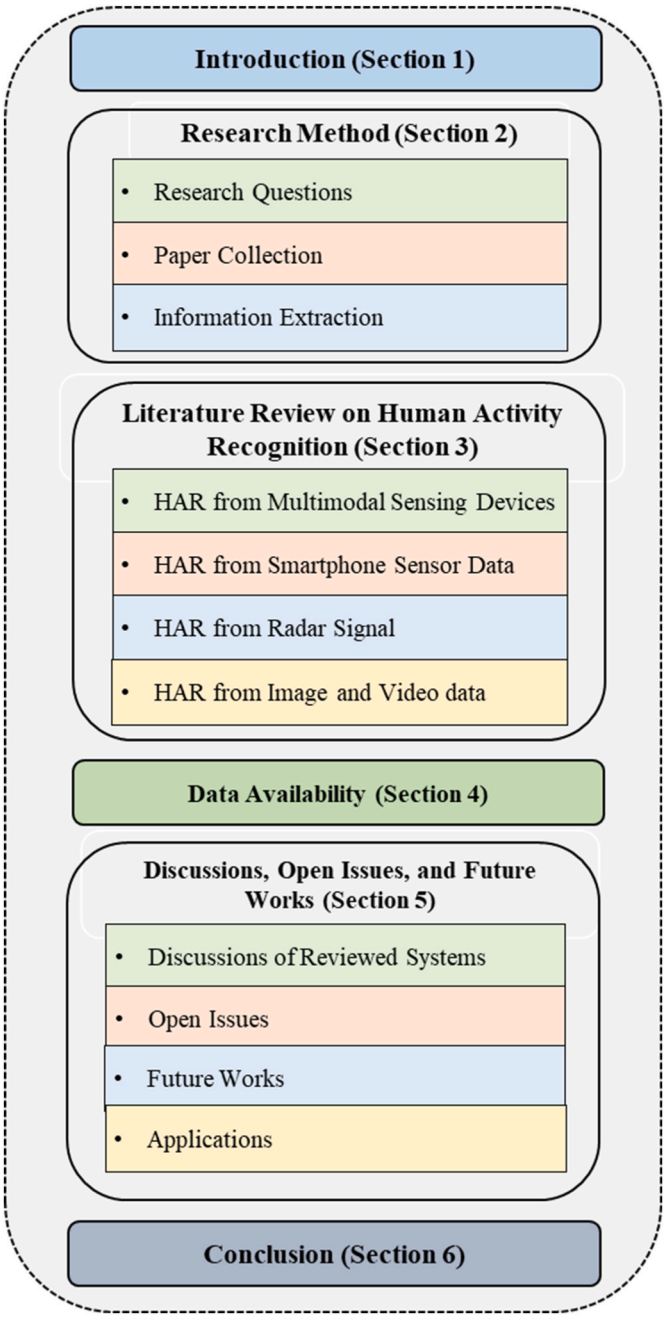

Fig. 2. Organization of our systematic review.

of human activity recognition platforms using the data from wearable sensors and smartphones. Moreover, this review provided an in-depth analysis of the available benchmark dataset for HAR. Dhillon et al. [36] discussed various deep learning models such as two-stream networks, C3D, and RNN for HAR. This survey considered the data from RGB cameras in the form of videos, depth maps, and skeleton points. Finally, quantitative analysis of the developed HAR systems is demonstrated in terms of accuracy. In another survey, Ramamurthy and Roy [37] presented the new trends of machine learning and data mining approaches for human activity recognition highlighting the fundamental problems and challenges. This survey categorized the reviewed systems based on the use of machine learning and data mining architectures and described the developed frameworks accordingly. The implementation details of the algorithms used for HAR are not mentioned in the review. Verma et al. [38] considered the surveillance system to demonstrate the progress of behavior recognition frameworks using supervised and unsupervised machine learning approaches. The review described various normal and abnormal human activities recognition along with different feature selectors and detections utilized in the existing literature. However, the survey lacks the description of the benchmark datasets used in the reviewed platforms. Biswal et al. [39] described various benchmark datasets for HAR focusing on the different levels of activities and the data acquisition techniques. This survey discussed the different machine learning approaches with implementation details. Although this review presented the potential future works, no open issues and challenges are described.

Further, Kumari et al. [40] reviewed the increasing trends of wearable devices and multimodal interfaces for HAR where they discussed the fundamental requirements, architectures, the current market scenario, and developments utilizing the concept of wearable sensors and biological signals. The authors of [41] reviewed the human activity recognition frameworks that are developed based on accelerometer data only. This review presented some features such as sample rate, window size, and percentage of overlap for time-series data to compare the existing literature. Additionally, the feature extraction techniques, as well as the factors which affect HAR are highlighted briefly in this review. Cornacchia et al. [42] presented the HAR systems based on the role of wearable sensors including pressure sensor, accelerometer, gyroscope, depth sensors, and hybrid modalities. The recent works of HAR are categorized based on how the sensor data is processed using machine learning approaches. Furthermore, Beddiar et al. [43] surveyed the recent progress of HAR highlighting some prominent features such as the type of activities, input data type, validation mean, body parts, and input viewpoint. A comparison was made among the state-of-the-art based on the type of activities. In addition, a detailed description of the vision datasets was highlighted in this review. Zhang et al. [26] conducted a review highlighting vision-based HAR that considered the data types and the learning methods. This review categorized the state-of-the-art HAR based on the use of types of images/videos such as color, skeleton, and data. Accuracy is used as the only performance measure to compare the existing HAR platforms. Bux et al. [44] reviewed the vision-based HAR depending on the use of various stages for recognition purposes such as object segmentation, feature extraction and representation, and activity recognition. However, this review only considered the videos and excluded the image data. To show the novelty of our review, a comparative analysis of existing state-of-the-art for human activity recognition is depicted in Table 1.

Although several reviews have been conducted for HAR based on deep learning and machine learning techniques, to the best of our knowledge, this is the first article that specifically highlights the convolutional neural network-based systems developed for HAR in the scope of multimodal sensors, smartphone, radar data, and vision data. The key contributions of the paper are demonstrated as follows.

(i) This work provides a comprehensive survey of CNN-based HAR that includes the recent progress as well as an in-depth analysis of the developed systems to serve the research community.
(ii) The survey discusses the developed systems in terms of performance, strength, weakness, and the used hyperparameters.
(iii) This work also provides a data availability perspective for developing HAR systems. General characteristics of available benchmark datasets as well as specifics such as number of types of HAR in each dataset, number of samples, number of sensing points, types of sensors used, and position of the sensors are presented.
(iv) An in-depth analysis is provided to discuss the systems that are best suited for handling all of the varied scenarios of HAR.
(v) Lastly, the challenges of the existing systems are highlighted while outlining possible suggestions and improvements for future research directions.

The structure of our systematic review is illustrated in Fig. 2. The rest

of the article is arranged as follows. The research methodology including research questions, paper selection, and information extraction process from a large number of existing literatures is described in section 2. Section 3 reviews the most recent works that are developed based on CNN architecture. A brief discussion on the available public datasets that are widely used in human activity recognition research is presented in Section 4. The current state of the art with open issues, suggestions for potential future work, and applications are discussed in Section 5. Section 6 concludes the paper.

2. Research method

As our goal is to provide a systematic review of CNN-based HAR systems–we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [45] methodology for streamlining the related paper selection process. The major sequential steps in the PRISMA process are as follows: identification of relevant papers, elimination of duplicate papers through scanning, filtering selection based on eligibility, and creation of an ultimate review list through the inclusion of selected papers. The paper selection process through PRISMA is illustrated in Fig. 3. We only considered the most recent works published in the last ten years in this field for this review work. We initially collected 463 papers from Web of Science, 162 papers from NCBI, 776 papers from Google Scholar, 46 papers from PubMed, and 43 papers from CINAHL through our initial search–resulting in 1490 total papers. The majority of the initial papers were collected from Google Scholar. The keyword "Convolutional Neural Network for Human Activity Recognition" resulted in the highest number of papers. We discarded 511 duplicate papers. We also excluded 440 irrelevant papers based on their topic. We fully read 276 papers after the quality assessment. The quality of the papers was assessed based on representation, plausibility, and methodological mistakes. The final review contains 63 papers. The present systematic review follows the three steps presented in Ref. [46]: definition of research questions, paper collection process, and information extraction. The following subsections expand on the three steps in the context of this study.

2.1. Research questions

We aim to present a systematic review of the modern CNN-based HAR systems for various data modalities. We provide a comprehensive survey of CNN-based HAR systems including the recent progress and in-depth analysis of the reviewed systems. We discuss the reviewed systems in terms of methodology, how they improve on previous systems, performance, strengths, weaknesses, and specific implementation details such as hyperparameters. We present and discuss the available HAR public benchmark datasets and their characteristics. The open issues challenges related to CNN-based HAR research and recommendations to overcome these challenges are also provided. We also highlight and suggest future research directions. We used the following research questions to guide the systematic review process in general and the article selection process in specific: "What is the state-of-the-art concerning the usage of CNN for HAR for various modalities?", "How did the CNN methods evolve to overcome the drawbacks in previous works?", "Which datasets are most commonly used for each task?", "How do the publicly available datasets differ from one another based on the used sensing modalities, sensor positions, sensing points, data collection scenario, and the number of subjects?", and "What common issues are still present in modern CNN based HAR systems and how can be they overcome?".

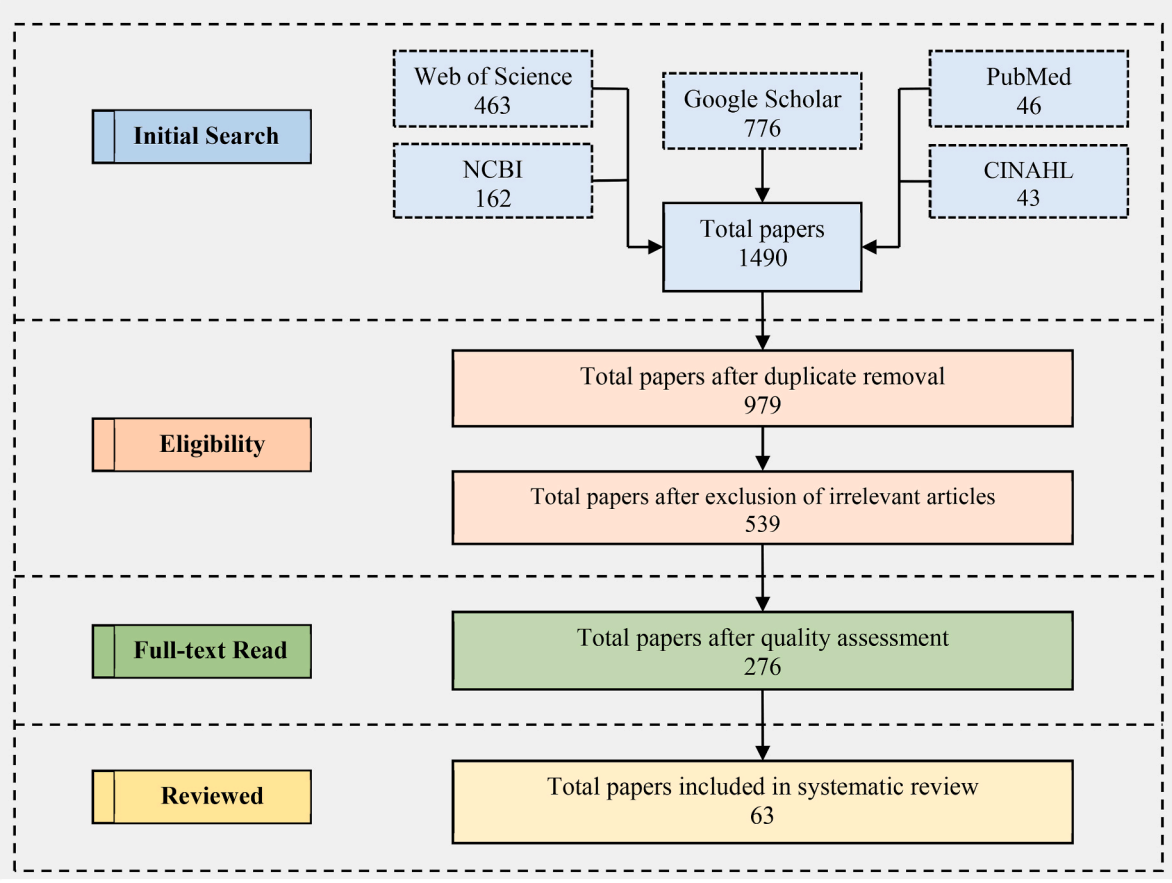

Fig. 3. The flow of article selection process for final review

2.2. Paper collection

We used the following sources to search and collect relevant papers: Google Scholar, PubMed, NCBI, CINAHL, and Web of Science. We used the following keywords for manual web searches and for searching the databases: "CNN for human activity recognition", "radar-based human activity recognition using CNN", "vision-based human activity recognition using CNN, "CNN for multi-modal human activity recognition", "CNN for smartphone-based human activity recognition", "public datasets for human activity recognition", "vision datasets for human activity recognition", "benchmark datasets for human activity recognition, depth image datasets for human activity recognition, and inertial datasets for human activity recognition". We analyzed the titles and abstracts of the research works found through the search results to eliminate duplicates and to ensure the selected articles fall within the scope of this review work. The specific topics of interest were identified and quantified after thoroughly analyzing each selected article. During the quality assessment process, some articles were removed from the review due to reasons such as: using machine learning algorithms including boosting, random forest, decision tree, and support vector machine (SVM), and using enterprise auto-machine learning solutions. To create a bibliography, we used Mendeley as the citation management tool. The broad inclusion requirements were such that all papers: should be written in English, should utilize CNN for feature extraction or classification purposes, and should be related to the field of HAR.

2.3. Information extraction

We extracted the following information from the reviewed papers: the datasets that were used, data preparation strategies, sensing modalities, feature selection and extraction process if available, specifics of used models, how the methodology improves upon previous studies, performance, strengths, and weaknesses. For the reviewed data sources, we extracted the following information: author information, publication year, number of subjects, sensing scenario, spontaneity, types of HAR, number of samples, sensing points, used sensors, and position of the sensing devices.

3. Literature review on human activity recognition

With the rapid growth and performance improvement of deep learning techniques particularly CNN architectures, numerous researchers have adopted them for dealing with HAR problems. This paper is focused on reviewing the four commonly used categories of devices for HAR: i) Human activity recognition from multimodal sensing devices, ii) Human activity recognition from smartphone sensor data, and iii) Human activity recognition from radar signals, and iv) Human activity recognition from image and video signals - as well as the CNN tools that have been used in conjunction with these devices for performing activity recognition. Fig. 4 illustrates the human activity recognition systems and their different types based on the sensing devices, signals, or sensing modalities. The multimodal sensing signals based human activity recognition systems generally use data from the following sensing modalities: accelerometer, gyroscope, magnetometer, and barometer. Smartphone-based human activity recognition systems utilize smartphone sensors to collect and classify data. Radar-based human activity recognition systems use various types of radars to collect data. Popular radar variants include Doppler radar, Frequency-Modulated Continuous Wave (FMCW) radar, interferometry radar, and Ultra-wideband (UWB). Vision-based systems collect data through RGB cameras or RGB-Depth (RGB-D) cameras. The recent works for human activity recognition are described below where we discuss the data collection strategies of each modality and the developed systems for each modality subsequently. In the next few sections, we discuss the four categories of devices and how the CNN-based models and tools are utilized to infer human activity from captured data using those devices.

3.1. HAR from multimodal sensing devices

In recent years, the research for activity recognition is focused on the combination of multiple sensor data (accelerometer, and gyroscope) that may increase the performance of the developed system in certain cases [47,48].

We consider various embedded body-worn solutions such as smart watches, smart necklaces, bands, helmets, and watches containing sensors like 3-D accelerometer, gyroscope, magnetometer, and barometer excluding smartphones as multimodal sensing devices [49]. These solutions are generally smaller than modern smartphones and require less power to run. However, these solutions oftentimes do not have Global Positioning System (GPS) and network connectivity. These devices generally do not require everyday charging. As these devices are always worn or kept on a person by the users, the sensors can record human motion and process them. Human activity generates acceleration and angular velocity. 3-D accelerometer sensor can sense acceleration along the 3-axes. A gyroscope and magnetometer can be used to sense the angular velocity and orientation. A barometer helps in sensing height changes during activities. Combining these properties, multimodal sensing solutions can infer human activities.

Wearable sensors have become increasingly popular in a wide range of applications as they can provide accurate and reliable data on daily

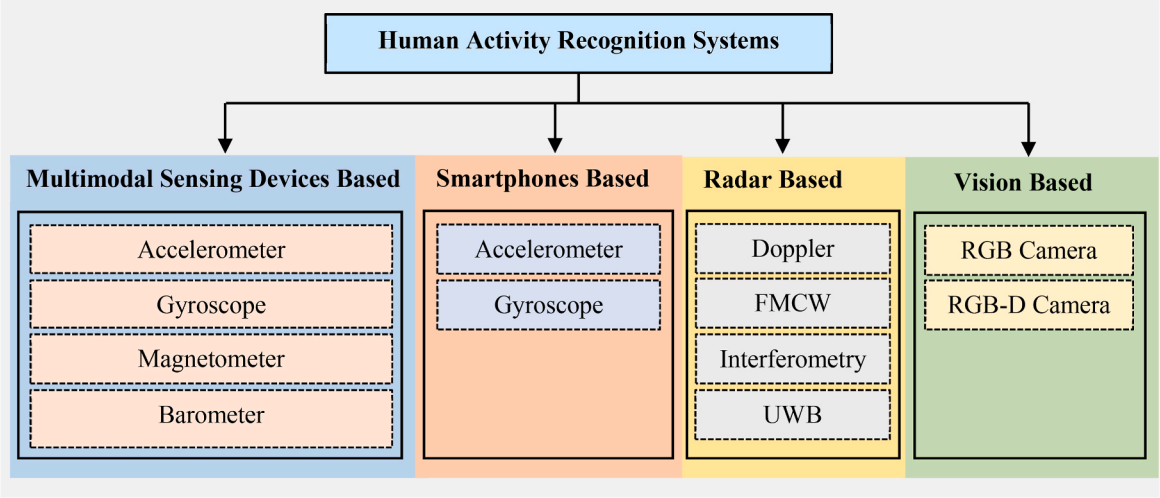

Fig. 4. Human activity recognition systems and their modalities.

human activities to ensure an ambient assisted living environment for the elderly [50]. Wearable sensors are capable of perceiving human body movements directly and efficiently over a long duration. The rapid growth of wearable technology has enabled the development of several types of smart sensors to capture physiological parameters accurately with lower power consumption and fewer processing resources. These sensors can be easily integrated into smartphones, bands, watches, and even clothes. A brief description of a few sensors that are commonly used for HAR is given below.

The most popular and commonly used sensor for HAR is an accelerometer [51]. An accelerometer is a sensor utilized to determine the acceleration - which is the rate of change of the velocity of an object. The sampling frequency of an accelerometer typically falls between tens and hundreds of Hz. There are numerous types of accelerometers on the market that use variable capacitance, piezoresistive, or piezoelectric transduction techniques. The operating concept of almost all types of accelerometers is the same: a mass reacts to acceleration by forcing a spring or an equivalent component to expand or compress according to the measured acceleration. A gyroscope [52] is also a frequently used sensor along with the accelerometer and mounted on the same body parts for human activity recognition. A gyroscope is a sensor for measuring angular velocity and orientation. Similar to an accelerometer, the sampling rate of a gyroscope sensor ranges from tens to hundreds of Hz. As a gyroscope has three axes, it also offers three separate time sequences. Another commonly used wearable sensor for HAR is a magnetometer [53] that is integrated with an accelerometer and a gyroscope into an inertial measurement unit. This sensor determines the change of a magnetic field at a specific position. The sampling frequency of this sensor is similar (tens to hundreds of Hz) to the accelerometer and gyroscope. A magnetometer sensor has three axes as well.

Table 2
Depth insight of the reviewed systems highlighting used method, performances, strengths, and weaknesses for multimodal sensing devices.

Authors	Year	Dataset	Used Method	Performances (%)	Strengths	Weaknesses
Ha et al. [56]	2015	mhealth [75], skoda [76]	2-D CNN	Accuracy = 98.29	The developed framework can capture local and spatial dependency using 2D kernels.	The proposed system acquired a slight temporal pattern.
Yang et al. [63]	2015	OPPORTUNITY [77], hand gesture dataset [78]	2-D CNN	Accuracy = 96, F1 Score = 95.5	An efficient illustration of the local salience of the raw data.	The testing time of the developed system is comparatively high (8 min approximately).
Jiang et al. [57]	2015	UCI [79], USC-HAD [80], SHO [81]	DCNN+, DCNN	Accuracy = 99.93	The framework can learn low-level to high-level features.	The computation cost of the developed scheme is comparatively high as the size of the input matrix rises gradually.
Ha et al. [62]	2016	mhealth [75]	2-D CNN	Accuracy = 91.9	Learned modality-specific features through partial and full weight sharing in the lower layers.	Many attempts are being made to achieve minor improvements of the state-of-the-art.
Choi et al. [58]	2019	Gas sensor arrays dataset [82], OPPORTUNITY [77]	Modified CNN (EmbraceNet)	F1 Score = 91.2	The use of docking and embracing layers improves the performance of the developed system.	The system cannot adopt rapidly changing sensors combinations.
Lawal et al. [61]	2020	RWHAR [83]	2-D CNN	Precision = 90, Recall = 90, F1 Score = 90	Recognized the sensor's position as well as human activities.	The efficiency of the framework is slightly low for practical use.
Yen et al. [60]	2020	UCI open dataset [84], and recorded data in this work.	1-D CNN	Accuracy = 95.99, Precision = 96.04, Recall = 95.98, F1 Score = 96.01	The developed system achieved comparatively good performances for all evaluation metrics.	The prototype has not been fully ready yet for application purposes.
Choi et al. [59]	2018	Real-time data collected from the participants.	CNN + RNN	F1 Score = 68	Considered the importance of each modality data based on confidence scores.	The developed scheme somewhat obtained low performance.
Andrade-Ambriz et al. [71]	2022	KARD dataset [85], CAD-60 dataset [86], MSR daily activity dataset [87]	3-D CNN + LSTM	Precision = 100, Recall = 100	The architecture combines time-motion attributes with the spatial location of human activities.	The training and classification time is a little bit high for real-time uses.
Wang et al. [67]	2019	UCI [79], Weakly labeled dataset	2-D CNN + Attention	Accuracy = 93.83	The system can recognize long sequence activity.	The scheme failed to recognize multiple types of activity.
Zhu et al. [72]	2019	UCF101 [88], HMDB51 [89]	Spatial CNN, Temporal CNN, and Attention	Accuracy = 94.94	The framework can retrieve significant features for recognition across multiple perspectives.	The effects of variation of data in different modes are not demonstrated here.
Gao et al. [73]	2020	WISDM [90], OPPORTUNITY [77], Weakly labeled dataset	Residual network + Dual attention	Accuracy = 98.85	The proposed attention mechanism can extract the most crucial data from a long sensor sequence, increasing the comprehensibility of the sensor signal.	The framework combines insignificant sensor modalities, which introduces significant noise.
Hamad et al. [66]	2021	UCI [79], Ordonez smart home dataset [91], Kasteren smart home dataset [92]	Dilated causal convolution + multi-head self-attention	F1 Score = 92.24	The system combines multi-scale contextual data to create insightful feature space.	The operation self-attention mechanism grows quadratically with the input signal, causing training time to increase.
Tan et al. [68]	2021	CAD-60 dataset [86]	CNN + attention-based LSTM	Precision = 97.02, Recall = 96.83	The proposed network can extract posture features from skeleton joints from RGB images efficiently.	The system failed to detect complex activities.
Tang et al. [74]	2022	WISDM [90], UNIMIB SHAR [93], PAMAP2 [94], UCI [79], Weakly labeled dataset	ResNet + Attention	F1 Score = 98.61	The triplet attention obtains a high level of generalization ability.	The developed scheme has a large number of parameters that increased the complexity.

*Note: Some reviewed systems use multiple datasets in their experiments. Here, we considered only the best performance.

Electrocardiography (ECG) [54] is a biometric sensor for HAR that detects the electrical signals produced by the heart. The information about the rate and regularity of heartbeats is extremely useful provided by the ECG signal. It is challenging to analyze subject variations in the ECG data because everyone's hearts vibrate in their unique ways. The output of an ECG sensor is univariate time series data. Electromyography (EMG) [55] sensor is used for human activity monitoring that detects muscle response or electrical signal in response to a nerve's stimulation of the muscle. Both the EMG and ECG sensors are required to be attached to human skin to record the data. Even though EMG is less frequently employed in traditional situations, it is more suited for fine-grained motions including hand or arm movements, and facial expressions. The EMG sensor generates the output in the shape of univariate time series.

In multimodal sensing devices, a big challenge is learning the inter-modality correlations along with the intra-modality data for CNN-based human activity recognition. To solve this issue, some CNN-based approaches have been developed that combine various modalities for the development of single extracted features or ensemble the output of different architectures. The simplest way to handle multimodal sensor data is combining the data from all sensors ignoring the sensor modalities although it has a chance of losing accurate correlations. In order to capture local and spatial dependence over time and sensors respectively, a multi-modal CNN architecture is proposed in Ref. [56] that used 2-D kernels in both convolution and pooling layers. As the relationship between the non-adjacent modalities is absent from traditional CNN, the system developed in Ref. [57], proposes a novel architecture in which any sequence of signals can be adjacent to every other sequence. Several systems have already been developed that consider each sensing modality initially and then combine them. This architecture provides modality-specific information along with versatile distribution of complexity. An architecture named EmbraceNet was proposed in Ref. [58] that processes the sensors' data separately and feeds them into the EmbraceNet. An effective way of fusing multimodal information using this architecture is through docking and embracing structure. A deep multimodal fusion architecture is introduced in Ref. [59] that calculates the confidence score of each sensor automatically and combines the features of multi-modal sensors based on the scores.

Some of the frameworks are introduced to minimize the interference between the used sensors by treating each sensor axis independently. In this case, 1-D CNNs are very popular for feature learning of each separate channel. A 1-D CNN architecture, proposed in Ref. [60] that omitted the pooling layers to achieve more detailed features where the data from the accelerometer and gyroscope are fed into the network separately. In Ref. [61], the acceleration of the accelerometer and gyroscope from seven different body positions are used to produce frequency images that are served as the input of two-stream CNN to learn inter-modality features. Very recently, a few CNN-based systems have been developed to handle univariate multichannel time-series data where the same sized filter is implemented to all time sequences. In this scenario, the raw signal is converted to a 2-D array by stacking along the modality axis which is then applied to 2-D CNN with 1-D filters [62,63]. However, the characteristics of all sensor data do not combine externally, but they interact through mutual 1-D filters.

There is another new trend in the area of deep learning called attention mechanisms [64,65] that has become a very popular and frequently used concept in diverse application domains including human activity recognition in recent years. In the current scenarios, the majority of the developed systems used shallow feature learning architectures that could not recognize human activities accurately in real-world situations. To solve this issue, Hamad et al. [66] used a dilated causal convolution with multi-head self-attention for physical activity recognition. During recognition, the multi-headed self-attention is utilized to allow the model to highlight significant and vital time steps rather than irrelevant time steps from the sequential feature space. The proposed architecture obtained an F1 Score of 92.24% from the experimental findings. Wang et al. [67] introduced an activity recognition system that processed weakly labeled information utilizing attention mechanisms. The compatibility between global features and local features is computed using this approach. By weighing their compatibility, the attention mechanism enhanced the salient activity data and suppressed the insignificant and slightly confusing data. The experimental results revealed that the scheme appraised an accuracy of 93.83%. Tan et al. [68] presented faster region-based CNN and attention-based LSTM for human activity recognition where the CNN extracted the feature as posture vector and the BiLSTM architecture classified the human activities. An attention layer is added between the two BiLSTMs. The developed network obtained precision of 97.02% and recall of 96.83% from the experimental findings.

In the current state of the art, most of the existing frameworks for HAR take the global information of input sequence and avoid local information that demonstrates changes in behaviors, causing the method to be responsive to external factors including occlusion and illumination change. To resolve this problem, some of the studies [69,70] consider the local spatial features, global spatial features, and temporal features for HAR. Andrade-Ambriz et al. [71] proposed a human activity recognition framework using a temporal convolutional network (TCN) that utilized spatio-temporal features as input of the architecture. The experiment demonstrates that the developed prototype achieved 100% precision and recall for two popular datasets. The scheme shared the activity recognition results to a robot called NAO during real-world environment testing. Zhu et al. [72] introduced a multimodal activity recognition scheme that fused three spatial features: local, global, and temporal features of input signals to classify different human actions. The proposed system divided the input into three segments where the global spatial features are found from the first segment (RGB frame) using a spatial CNN, the local spatial features are extracted from the local blocks utilizing another spatial CNN, and the last segment (optical flow) is utilized to extract temporal features through the use of a temporal CNN. The three architectures are evaluated individually using two benchmark datasets and the final output is obtained using the weighted sum of the three networks. The best accuracy of 94.94% is found from the experimental results for the UCF101 dataset. Gao et al. [73] proposed a framework called DanHAR that combined channel and temporal attention on residual networks to enhance feature representation capability for human activity recognition. The proposed architecture takes a time window as input and sends it to convolutional layers to obtain visual features. This network then generates channel attention through pooling layers (max-pooling and average-pooling) to combine features along the temporal axis. It is found from the experimental results that the proposed architecture obtained an accuracy of 98.85% for the WISDM dataset. In another research, Tang et al. [74] developed a triplet cross-dimension attention model for HAR, which introduced three attention parts to make the cross-interaction between sensor, temporal, and channel dimensions. The performance measure shows that the F1 Score of 98.61% is obtained by the developed system. It is worth mentioning that the system is tested in a real-time environment using a Raspberry Pi prototype.

Table 2 illustrates the in-depth analysis of the reviewed systems based on the used techniques, performances, strengths, and weaknesses for the modality of multimodal sensing devices.

3.2. HAR from smartphone sensor data

Smartphone has become very popular for HAR thanks to the rapid growth of modern technology as it has various built-in sensors for this task [95]. The major problem of the traditional wearable sensing device is that the users should carry an extra device; sometimes they are not willing to carry it, or a few times get forgotten. As almost everyone now has a smartphone, it has become an excellent choice to conduct research using smartphone sensor data, which ensures the portability of the developed systems to a great extent [96].

Table 3
Depth insight of the reviewed systems highlighting used method, performances, strengths, and weaknesses for smartphone sensor data.

Authors	Year	Dataset	Used Method	Performances (%)	Strengths	Weaknesses
Ravi et al. [101]	2016	ActiveMiles [101], WISDM v1.1 [90], Daphnet FoG [107], skoda [76]	Temporal CNN	Accuracy = 98.2	The novel features generation method is the sum of the temporal convolutions of the transformed input.	As for resource limitations and the simple strategy of this system, the performance cannot outperform the shallow features-based frameworks in some cases.
Almaslukh et al. [98]	2018	RealWorld HAR [108]	2-D CNN	Precision = 89, Recall = 87, F1 Score = 88	No generality losses have happened.	Only statistical attributes are considered instead of position-independent handcrafted attributes.
Nair et al. [106]	2018	UCI [79]	Encoder-Decoder TCN, Dilated TCN	Accuracy = 97.8, F1 Score = 97.7	The raw sensor data is used instead of the more expensive pre-processing.	In complex scenarios, the built-in sensors in smartphones are unable to collect a large number of accurate data.
Lee et al. [99]	2017	$Feature_{10}$, and $Feature_{20}$ dataset	1-D CNN	Accuracy = 92.71	Minimized the potential rotational interference of the raw data.	Correlation losses occur due to the transformation of vector magnitude from raw signals.
Ravi et al. [100]	2017	ActiveMiles [101], WISDM v1.1 [90], WISDM v2.0 [109], Daphnet FoG [107], skoda [76]	1-D CNN	Accuracy = 98.6	The inertial sensor information along with complementary data are learned from shallow attributes.	Although the developed system is recommended for real-time application, the latency is slightly high.
Nafea et al. [105]	2021	WISDM [90], UCI [79]	CNN + BiLSTM	Accuracy = 98.53, Cohens Kappa = 98	The relationship between the movement and spatial features is maintained effectively here.	The framework has not experimented with actual users in real-time environment.
Zhang et al. [103]	2020	WISDM [90]	CNN + Attention	Accuracy = 96.4, F1 Score = 95.4	Manual feature extraction is not required. It learns the necessary features automatically.	The complexity of the developed system is relatively high.
Ge Zheng [104]	2021	WISDM [90], UCI [79]	2-D CNN + Attention	F1 Score = 95.69	Global spatial-temporal features are learned efficiently.	The features in the space domain are not handled here.
Khan and Ahmad [102]	2021	WISDM [90], UCI [79]	1-D CNN + Attention	Accuracy = 98.18, Precision = 97.12, Recall = 97.29, F1 Score = 97.20	Squeeze-and-excitation module enhanced the performance of the lightweight architecture.	The developed framework cannot recognize complex activities properly.

*Note: Some reviewed systems use multiple datasets in their experiments. Here, we considered only the best performance.

All modern smartphones contain sensors such as accelerometer, gyroscope, and magnetometer. Smartphones generally require more power than multimodal sensing solutions. However, almost all smartphones require regular daily recharging. Smartphones have more processing powers than multimodal embedded solutions. Smartphones also have GPS and network connectivity, making them viable for transferring data and decisions in various client-server models [97]. Thus, smartphones have access to more sophisticated data-heavy models. Users generally carry smartphones on locations such as thigh, chest, and hand. As human motion generates acceleration and angular velocity, smartphone

Table 4
Depth insight of the reviewed systems highlighting used method, performances, strengths, and weaknesses for radar signals.

Authors	Year	Dataset	Used Method	Performances (%)	Strengths	Weaknesses
Ye et al. [113]	2020	Data collected using Infineon Sense2GoL Doppler radar.	Fourier CNN	Accuracy = 98.6	The use of Fourier layer, dilated convolutions, and triplet loss increased the performance.	The micro-Doppler effects somewhat increased the feature extraction complexity.
Ye et al. [112]	2019	Real-time data collected for this study.	1-D CNN	Accuracy = 96.31, Precision = 97, Recall = 97, F1 Score = 97	The system can extract higher-level hidden attributes.	The balance between performance and power requirements is not mentioned.
Chen et al. [114]	2020	Real-time data collected from the applicants.	1-D CNN	Accuracy = 96.1	The complexity of the developed system is comparatively low.	The developed system is somewhat affected by environmental interference.
Erol et al. [116]	2019	A 25 GHz software-defined radar system is used to collect the data	ACGAN-DCNN	Accuracy = 82.56	The system is capable of a few numbers of samples by using augmentation.	The performance of the framework is not well enough for practical uses.
Alnujaim et al. [115]	2020	Data collected from the environment using Doppler radar.	GANs + DCNN	Accuracy = 99.55	The system showed high performances with synthetically generated data trained with GANs.	The training of multiple GANs would be slightly harder with the increasing number of activities.
Alnujaim et al. [118]	2021	MOCAP [119]	GAN + U-Net	NMSE = 0.50E-04	This study ensures the diversity of the dataset by augmenting radar data from various aspect angles.	The developed scheme generated data samples with a high degree of similarity due to the convergence instability of GAN.
Wu and Ye [117]	2021	A Doppler radar is used to collect seven human activities.	GAN + CVAE	Accuracy = 94.89	The use of unsupervised pre-training and adversarial training fits the system to obtain better accuracy.	No experimental setup is shown to retrieve the data samples.

*Note: Some reviewed systems use multiple datasets in their experiments. Here, we considered only the best performance.

applications can sense these changes through the embedded smartphone sensors and process them. Smartphones also provide high-level access to sensor data via the Software Development Kit (SDK) of the operating systems. SDK also provides additional support such as always-on applications, notification and alarm systems, and sensor data buffers.

One of the major issues that should be considered is the position of the smartphone as it can be kept in a pant pocket, hands, bags, and shirt pocket. It is evident that due to the various locations of the smartphone, the raw signals change considerably as the movements of the various parts of the body are different. To handle this problem, some of the reviewed literature developed position-independent solutions. A smartphone-based position-independent activity recognition system using CNN was introduced in Ref. [98] that used time-domain statistical features. Here, mean-centering is used to convert the raw input to an appropriate form to train the optimum threshold without any bias. A 1-D CNN is introduced in Ref. [99] that used smartphone accelerometer data from different body positions like the bag, hand, and pocket for activity recognition. As the data for different body positions are used, the developed system effectively ensures the position-independent property.

Another big challenge for smartphone-based systems is that traditional deep learning architectures cannot be easily embedded in such systems due to low power capacity and limited computational capacity. To resolve this issue, a few of the systems have been developed that merged the hand-crafted features and deep features for activity recognition. Decreasing filter size is a potential solution to reduce the size of the network that optimizes computing operations. A HAR system is introduced in Ref. [100] where the deep features and hand-crafted features are arranged in parallel, and the features are then incorporated into the 1-D CNN architecture. The performance of the developed system has been increased with a small number of computational operations. The system developed in Ref. [101] used just one CNN layer and two fully connected layers where the spectrogram features are fed into the network for activity recognition. The experimental findings revealed that the system achieved milliseconds to tens of milliseconds of computing time for a single prediction.

In another study, the authors of [102] proposed an attention-based multi-head architecture for HAR. The developed architecture had three lightweight convolutional heads; each is designed to extract features from collected data using 1-D CNN. The lightweight multi-modal architecture is stimulated with an attention mechanism to improve CNN's representation capability, enabling the automatic selection of significant features while suppressing irrelevant ones. Although two datasets have been utilized here, the highest accuracy, precision, recall, and F1 Score of 98.18%, 97.12%, 97.29%, and 97.20% respectively are achieved from WISDM data. Zhang et al. [103] developed a system that combined the concept of CNN and attention mechanism for activity recognition using the data from a smartphone. Here, the attention is incorporated into multi-head CNNs that facilitate extracting and selecting features efficiently. The proposed scheme achieved accuracy and F1 Score of 96.4% and 95.4% from the experiments. The author of [104] introduced a novel deep learning architecture called LGSTNet, combining the concept of CNN and attention mechanism to recognize human activity from the data of accelerometers and gyroscopes. The activity window is fragmented into various sub-windows in this system, and the local spatial-temporal attributes from those sub-windows are learned using an attention mechanism and CNN. It is evident from the experiments that the obtained F1 Score of the proposed network is 95.69%.

In another research, Nafea et al. [105] presented a HAR framework that used spatial information and temporal information, extracted by CNN with differing kernel sizes and Bi-directional Long Short-Term Memory (BiLSTM), respectively. The retrieved spatio-temporal data were merged in a mixed model that was trained and verified using two benchmark datasets, yielding a 98.53% accuracy, and Cohens Kappa of 98% for the WISDM dataset. Nair et al. [106] developed a temporal convolution network-based network to recognize human activities from raw motion data collected utilizing smartphone sensors. To deal with sequence information with big receptive fields and temporality, dilations and causal convolutions have been developed in this system. The accuracy and F1 Scores of 97.8% and 97.7% respectively are achieved from the experimental results using the encoder-decoder temporal convolutional network.

Table 3 depicts an in-depth analysis of the reviewed systems based on the approaches used, their performances, strengths, and shortcomings for smartphone sensor data.

3.3. HAR from radar signal

The current research is focused on the device-free approach as it does not include any devices to take while participating in any activities. To ensure a device-free solution, a radar-based system shows the best performance due to its insensitivity to daylight and environmental effects as well as contactless-manner [110].

Radar signal-based sensing modality is widely used for stationary surveillance. Radar or radio detection and ranging systems detect both living and inanimate objects through reflection [111]. Radar systems generate intermittent high-frequency radio waves and transmit them to the environment around them. Radio waves are electromagnetic waves that travel at the speed of light. After hitting objects in the environment, radio waves bounce off or reflect from them. Radio systems can receive the reflected radio signal and extract properties of the object including size, shape, distance, and movement from the time required to detect the reflection and the change in frequency due to collision. Radio signal-based sensing modalities deployed in surveillance situations can thus detect stationary objects, humans as well as human motions through the characteristics of the received signal and infer the human activity.

In general, the radar signal is converted to the time-frequency domain which is a separate part from the learning architecture that sometimes does not extract the optimal features. In some cases, the raw signal is transferred to the short-time Fourier transform (STFT) or 2-D matrix and the deep learning architecture treats the 2-D matrix as an optical image. However, the optical image pixels have high spatial correlations, whereas the 2-D radar matrix pixels have a lot of temporal correlations. Hence, treating them as the same is not optimal for classification purposes. To improve the performance of the radar-based systems, some researchers focused on the use of variants of CNN rather than conventional CNN to resolve these issues. An end-to-end network named RadarNet is introduced in Ref. [112] where the STFT is substituted by two 1-D convolutional layers. The developed system merged all the steps (micro-Doppler radar data representation, extraction of features, and classification) of HAR in a single architecture. F-ConvNet, another end-to-end architecture is proposed in Ref. [113] that used three convolutional layers for multi-scale feature extraction. A novel layer named Fourier layer is proposed here that includes Fourier initializations and two branches of processing for learning the real and imaginary segments individually. In addition, to improve the classification accuracy, dilated convolutions are used. To reduce the computation complexity, an end-to-end network (1-D CNN) is designed in Ref. [114] for activity recognition using radar signals. The proposed system used the inception densely block (ID-Block) that is customized for the proposed 1-D CNN where ID-Block is comprised of an inception module, network-in-network methods, and a dense network.

To solve the issue of limited training data, generative adversarial networks (GANs) are frequently utilized in recent times. In Ref. [115], a GAN is developed for HAR using micro-Doppler signatures of radar. While the GAN is trained with the original micro-Doppler images, it generates a lot of similar images like the original that are fed into traditional CNN for training. The use of the increased number of images enhances the performance of the developed system. Erol et al. [116] proposed a human activity recognition platform that used synthetic radar data generated by GANs. The system introduced an auxiliary

classifier generative adversarial network (ACGANs) to generate a large number of training samples. The average test accuracy of 82.56% is found from the experimental findings using the proposed GAN networks with DCNN architecture. The main purpose of the developed framework is to reduce the classification confusion among similar activities by increasing the size of the training dataset. The authors of [117] introduced a framework to use the discriminator of the GAN for human activity recognition using limited radar data. To initialize the parameter of the GAN, a convolutional variational autoencoder (CVAE) is utilized in the proposed system. From the experimental study, it is found that the scheme obtained an accuracy of 94.89% from a few numbers of data samples. In another work, Alnujaim et al. [118] used GAN to increase the time-frequency images retrieved from a single angle into images from numerous angles resulting in the improvement of the sample set from multiple viewpoints. The average normalized mean square error (NMSE) of the developed prototype is 0.50E-04. However, the developed framework generated data samples with high similarity due to the convergence instability of GAN.

Table 4 depicts an in-depth analysis of the reviewed systems based on the techniques employed, performances, capabilities, and limitations for radar signals.

3.4. HAR from image and video data

Due to advances in technology, both RGB and RGB-D cameras are easily accessible and cost-efficient. Vision-based systems are effective for the surveillance of large regions in an effective manner.

Image and video sensing modalities are more accessible and easier to setup than radar sensing modalities [120]. Even cheap RGB cameras nowadays have night vision capability through Infrared (IR) sensing. As these systems are stationary solutions, very simple techniques such as background subtraction can be used to localize and monitor motion in the surveilled area. The detected motions can then be passed to CNN models to infer human activity. With the advancement of specialized processors for neural network processing, RGB solutions with built-in neural network processors and network connectivity are available. These systems can capture images or video sequences and process them locally using saved deep learning models in offline mode or can send data to servers running more sophisticated models in online mode.

Although previous research advances were focused on traditional vision-based algorithms, current advances in both deep learning algorithms and hardware enabled us to deploy highly effective deep learning algorithms alongside vision systems. Computer vision-based human activity recognition systems face several challenges such as interclass variation and intraclass similarity, diverse and complex backgrounds, multisubject interactions, group interactions, videos from long distances, and low-quality images. Intraclass variation arises from separate people performing the same action in their own ways. Interclass variation arises from the numerous types of activity we perform in our day to day lives. Vision-based image and video datasets also have various types of backgrounds. Backgrounds differ in lab scenarios as well as in real-world scenarios. Image and video data also suffer from inherent complexities such as pixilation, aliasing, light level differences, viewpoint variations, and occlusions [121]. Video-based human activity recognition datasets are more common than still image-based datasets, as activities are regarded more as a sequence of actions than a one-off scenario.

Most of the available human activity recognition datasets contain video sequences of Activities of Daily Livings (ADLs). The system in Ref. [122] introduced an abnormal human activity dataset containing

Table 5
Depth insight of the reviewed systems highlighting used method, performances, strengths, and weaknesses for video and image data.

Authors	Year	Dataset	Used Method	Performances (%)	Strengths	Weaknesses
Wang et al. [132]	2013	UCF101 [88], HMDB51 [89]	2-D CNN	Accuracy = 91.5 (UCF101), Accuracy = 65.9 (HMDB51)	Presents TDD as a robust feature descriptor combining both hand-crafted features and deep-learned features. Performance in the spatial domain is equal to or better than other approaches.	Performs slightly worse in temporal domain than other two-stream two 2-D CNN approaches.
Shinde et al. [124]	2019	LIRIS [135]	CNN (YOLO)	Accuracy = 88.37, Precision = 89.88, Recall = 88.08, F1 Score = 88.35	Only classified a small number of selected frames from every video to determine the action in that video, thus reducing time, cost, and computation requirements.	Does not work well in videos that have multiple types of action sequences.
Gul et al. [122]	2020	Real-time data collected using an RGB camera.	CNN (YOLO)	Accuracy = 96.8, Precision = 90.1, Recall = 88.4, F1 Score = 89.3	High-speed real-time abnormal human action recognition.	Works only when a single person is performing the task. Does not work well for group activities/small objects/overlapping objects.
Ji et al. [130]	2013	KTH [136], TRECVID'08 [137]	3-D CNN	Accuracy = 90.2, Precision = 78.24	Uses 3-D convolutions instead of two-stream 2-D CNN networks solution for taking both spatial and temporal domains of video into consideration. Does not require huge storage solutions.	Required more labeled data than unsupervised approaches which might become a concern in the case of huge datasets.
Simonyan et al. [125]	2014	UCF101 [88], HMDB51 [89]	CNN + SVM	Accuracy = 88	Uses a two-stream two 2-D CNN network that is trained on both spatial and temporal domains of video data. The temporal domain takes stacked multi-frame dense optical flow data as input.	There are huge storage requirements to store optical flow data even for small datasets despite taking preventive measures such as type conversion and compression. Big data storage and management requires for large datasets.
Basavaiah and Patil [134]	2020	Weizmann dataset [138], and KTH dataset [136]	SIFT, optical flow, CNN	Accuracy = 98.03 (Weizmann), Accuracy = 94.96 (KTH)	The computation time is relatively low as the dissimilar videos are avoided in the input through the distance similarity measure.	The background and feature similarity are not taken into consideration; thus increased the misclassification.
Serpush and Rezaei [133]	2021	UCF101 [88]	HOG, CNN and LSTM, k-NN	Accuracy = 93.80	The system achieves relatively better performance through the use of multiple techniques for preprocessing and classification.	Spatio-temporal information is not considered in the proposed research.

*Note: Some reviewed systems use multiple datasets in their experiments. Here, we considered only the best performance.

Table 6
General characteristics of the datasets.

Dataset	Author	Year	Number of subjects	Scenario	Spontaneity
WISDM v1.1	Kwapisz et al. [90]	2010	29	Out-of-lab real-world ADL.	Not spontaneous
Kasteren* SH	Kasteren et al. [92]	2011	3	Out-of-lab real-life ADL	Spontaneous
WISDM v2.0	Lockhart et al. [109]	2012	59	Out-of-lab real-world ADL.	Not spontaneous
UCF101	Soomro et al. [88]	2013	N/A	Out-of-lab real-life ADL.	Spontaneous
Ordóñez* SH	Ordóñez et al. [91]	2013	1	Out-of-lab real-life ADL	Spontaneous
YouTube Sports 1 M	Karpathy et al. [128]	2014	N/A	Out-of-lab real-life ADL.	Spontaneous
mhealth	Banos et al. [75]	2015	10	Out-of-lab real-life ADL.	Spontaneous
ActiveMiles	Ravi et al. [101]	2016	10	Out-of-lab real-world ADL.	Spontaneous
YouTube 8 M	Abu-El-Haija et al. [129]	2016	N/A	Out-of-lab real-life ADL.	Spontaneous
Daphnet FoG	Bachlin et al. [107]	2010	10	In laboratory ADL data.	Not spontaneous
USC-HAD	Zhang et al. [80]	2012	14	In laboratory ADL data.	Spontaneous
PAMAP2	Reiss et al. [94]	2012	18	In laboratory ADL data	Not Spontaneous
CAD-60	Sung et al. [86]	2012	4	In laboratory ADL data	Not Spontaneous
MSR DA	Wang et al. [87]	2012	10	In laboratory ADL data	Not Spontaneous
UCI	Anguita et al. [79]	2013	30	In laboratory ADL data.	Not spontaneous
SHO	Shoaib et al. [81]	2014	10	In laboratory ADL data.	Not spontaneous
LIRIS	Wolf et al. [135]	2014	21	In laboratory ADL data.	Not Spontaneous
KARD	Gagilo et al. [85]	2015	10	In laboratory ADL data	Not Spontaneous
NTU RGB + D	Shahroudy et al. [139]	2016	40	In laboratory ADL data.	Not Spontaneous
UniMiB SHAR	Micucci et al. [93]	2017	30	In laboratory ADL data	Not Spontaneous
PKU-MMD	Liu et al. [140]	2017	66	In laboratory ADL data.	Not Spontaneous
KTH	Schüldt et al. [136]	2004	25	Combination of both	Not Spontaneous
OPPORTUNITY	Roggen et al. [77]	2010	12	In laboratory non-ADL data.	Spontaneous
skoda	Zappi et al. [76]	2008	1	Activity of assembly line worker in car production scenario.	Not spontaneous
HMDB51	Kuehne et al. [89]	2011	N/A	Clips from action movies.	Not Spontaneous

*Note: N/A = Not Available, ADL = Activities of Daily Living, SH = Smart Home, DA = Daily Activity, Kasteren et al., and Ordóñez et al. have multiple datasets which were combined for their entries.

abnormal actions such as coughing, chest pain, faint, vomiting, and taking medication while also implementing a real-time high-speed recognition algorithm based on the You Only Look Once (YOLO) architecture [123]. However, this system only works in cases when a single human is monitored. The system cannot recognize group activities, overlapping objects, and small objects due to the spatial constraints of the YOLO backbone.

While performing human action recognition from video sequences, initial CNN-based research works processed all of the frames of a video for recognizing tasks represented in the video. However, this was inefficient due to the huge time, computation, and memory required to process all the frames. An alternative solution is proposed in Ref. [124] where only a selected number of frames of a video are classified instead of all the frames of a video. This drastically reduced the computation time and computation requirements, while also making the system real-time. In general, 30 frames from a video are selected for classification in a deterministic way based on the total number of frames in the video. These selected frames are classified to determine the represented actions and their confidence levels. These confidence levels and actions are averaged to determine the final action and confidence level of the entire video. However, this system works best when a video sequence contains a single action group. When a video contains multiple actions or action groups, determining a single action across an entire video becomes problematic. The CNN methods discussed till now only take the spatial domain characteristics of the videos into consideration. However, temporal domain characteristics can also be extracted from the videos which might act as discriminating features. A two-stream CNN for human action recognition is introduced in Ref. [125]. The spatial domain stream is trained on the individual frames of the videos. The temporal domain stream is trained on stacked multiple-frame dense optical flow. The dense optical flow in the temporal domain contains motion data of the objects. The two separate streams are combined by calculating stacked L2 normalized SoftMax scores as features. These features are classified using multi-class linear SVM [126]. The main issue with this method is the huge memory requirements to store all the optical flow data. To reduce the size of the saved data, float point data were converted to integer point data, and the saved data was compressed using JPEG [127]. Despite these measures, the saved optical flow data took huge memory spaces for even smaller datasets. For huge datasets such as the YouTube 1 M [128], YouTube 8 M [129] this method would require big data storage solutions. An alternate solution for taking the features from the temporal domain into consideration is presented in Ref. [130]. As opposed to training two separate 2-D CNN networks in their respective spatial and temporal streams, the system developed in Ref. [130] used 3-D convolutions to extract features from both the spatial and the temporal domain. In this way, a single 3-D CNN can be used in place of two-stream two CNN solutions. This also effectively solves the data storage issue of [125]. The performance of this model is comparable to the two-stream two CNN solutions. However, this model requires more labeled data than unsupervised counterparts, and thus, for large video human action datasets, accurate labeling poses a big challenge.

Handcrafted features such as the Histogram of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), Motion Boundary Histograms (MBH) have been historically used for human action recognition [131]. The framework developed in Ref. [132] introduces Trajectory-pooled Deep-convolutional Descriptors (TDD), a combination of handcrafted features and deep-extracted features for human activity recognition. The handcrafted features and the features extracted using deep 2-D CNN networks are aggregated based on trajectory constrained pooling. Spatio-temporal normalization and channel normalization are further used to transform the feature maps. The TDD features are highly discriminative and are learned automatically. However, this method performs slightly worse than two-stream 2-D CNN solutions. While the performance of TDD is excellent in the spatial domain, its performance is worse than or comparable to the performance of two-stream solutions in the temporal domain. The authors of [133] proposed a hierarchical approach for complex human recognition that used background subtraction and HOG for image preprocessing, deep learning architectures (CNN, and LSTM) for feature selection and to maintain the previous data, and Softmax-k-nearest neighbors (k-NN) for classifying the human activities. The proposed architecture has been evaluated using the UCF101 dataset that contains 101 complicated human activities. Extensive experiments revealed that the developed system achieved an accuracy of 93.80%. In another research, a fusion of scale invariant feature transform (SIFT) and optical flow method are

Table 7
Sensing modalities and relevant information on the datasets.

Dataset	Number of types of HAR	Number of Samples	Sensing Points	Sensors	Position of Sensor
WISDM v1.1	6	1,098,207	1	S (A)	Right Thigh or Left Thigh.
Kasteren SH	8	74,880	23	IR, F, R, P, T, Me	Toilet, Shower, Doors, Cabinets, Walls, Drawers.
WISDM v2.0	6	2,980,765	1	S (A)	Right Thigh or Left Thigh.
UCF101	101	13,320	1	C	Mixed.
Ordóñez SH	11	50,400	12	IR, F, R	Toilet, Doors, Cupboard, Walls.
YouTube Sports 1 M	487	11,33,158	1	C	Mixed.
mhealth	12	16,740	3	Sh (A, G, M, ECG)	Chest, Right Wrist, Left Ankle.
ActiveMiles	7	4,390,726	1	S (A, G)	Placement-invariant.
YouTube 8 M	4,716	80,00,000	1	C	Mixed.
Daphnet FoG	2	1,917,887	3	A	Trunk, Thigh, Ankle.
USC-HAD	12	10,570	1	MN (A, G, M)	Front Right Hip.
PAMAP2	18	N/A	3	IMU (A, G, M), H	Chest, Wrist, Ankle
CAD60	12	5440	1	D	Mixed
MSR DA	16	23797	1	C, D	Mixed
UCI	15	10,299	1	S (A, G)	Waist.
SHO	7	9,730	5	S (A, G, M, L)	Right Thigh, Left Thigh, Right Waist, Right Upper Arm, Right Wrist.
LIRIS	10	828	1	C, D	Mixed.
KARD	18	2160	1	C, D	Mixed
NTU RGB + D	60	56,880	1	C, D, IR	Mixed.
UniMiB SHAR	9	11,771	1	S (A, Mi)	Left Thigh, Right Thigh
PKU-MMD	51	1,076	1	C, D, IR	Mixed.
KTH	6	2,391	1	C	Mixed.
OPPORTUNITY	18	27,000	72	A, G, M, Mi, CS	All over the body and environment.
Skoda	10	701,440	19	A	Left Hand, Right hand.
HMDB51	51	7,000	1	C	Mixed.

*Note: Sh = Shimmer wireless sensor, A = 3-D Accelerometer, G = Gyroscope, M = Magnetometer, Mi = Microphone, S = Smartphones, MN = MotionNode, L = Linear Acceleration Sensors, CS = Commercial Sensors, C = RGB Camera, D = RGB Depth Camera, IR = Infrared Sensor, F = Float Sensor, R = Reed Switch, P = Pressure Sensor, T = Temperature Sensor, Me = Mercury Contacts, IMU = Inertial Measurement Unit, H = Heart Rate Monitor. Notation X (Y, Z) indicates the sensing modalities Y and Z are contained within X and X is the main data collection point.

exploited [134] to extract the shape, gradient, and orientation features from videos of human action datasets. Additionally, CNN is used to recognize human activities by training and testing the datasets. Two popular datasets, namely Weizmann dataset and KTH dataset are utilized to evaluate the performance of the developed framework and appraised an accuracy of 98.03%, and 94.96%, respectively.

Table 5 shows an in-depth analysis of the reviewed systems based on the approaches utilized, performances, strengths, and flaws for video and image data.

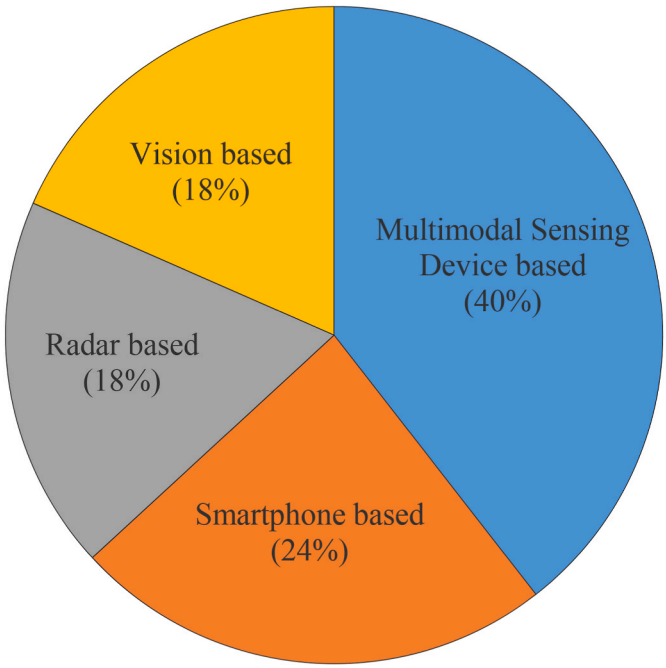

Fig. 5. Percentile representation of the modalities in the reviewed studies.

4. Data availability

In this section, we explore some of the common benchmark datasets for human activity recognition. The datasets were selected based on the frequency of their usage in the reviewed human activity recognition research works. Most of these datasets are numerical in nature. The datasets contain raw sensor data in some cases and transformed or fusion sensor data in other cases. In almost all cases, the number of test subjects is greater than 10. The data is also collected in different scenarios. In some datasets, the data was collected in laboratory environments. In other cases, the data was collected in out-of-laboratory real-life environments where the data was collected during the real-life events of the subjects. Some datasets are spontaneous, meaning in those cases the subjects were provided with the freedom to perform the activities in their own style. In other datasets, the activities and the steps to perform them were strictly maintained. The general characteristics of available datasets are depicted in Table 6.

Table 7 provides relevant information related to the datasets such as the number of types of HAR, number of samples, sensing points, the actual sensing modalities, and the position of the sensors. Almost all of the datasets collected sensor data for common activities of daily life such as standing still, jogging, running, jumping, cycling, crouching, climbing stairs, lying down, and walking. The PAMAP2 dataset [94] and the MSR Daily Activity dataset [87] contain some activities that are not generally present in the ADL dataset. Some of the exclusive activities represented in the PAMAP2 dataset [94] are cleaning the house, folding laundry, watching TV, and playing soccer. Some of the exclusive activities presented in the MSR dataset [87] are using a vacuum cleaner, cheering up, tossing paper, and playing guitar. The Kasteren et al. [92] smart home dataset contains activity records during users' using the toilet and shower. The Ordóñez et al. [91] smart home dataset contains activity records during users snacking, sitting idly on a desk, and grooming. The greatest number of sensing points was used in the OPPORTUNITY dataset. Various sensing modalities have been used to collect relevant data. The most commonly used sensing modalities are accelerometer, gyroscope, and magnetometer. The smart home datasets Ordóñez et al. [91] and Kasteren et al. [92] used rare sensors such as reed switches to record usage of doors and cabinets, float sensors to record toilet flushing, and passive IR sensors to record motion. A lot of the datasets used

Table 8
Hyperparameters of the CNN structures used in the reviewed systems.

Authors	Learning Rate	Number of Epochs	Batch Size	Dropout	Optimizer	Loss
Ha et al. [56]	N/A	N/A	60	0.5	N/A	N/A
Choi et al. [58]	0.001	500000	64	0.5	Adam	N/A
Yen et al. [60]	0.00001	600	N/A	0.5	Adam	Categorical_crossentropy
Lawal et al. [61]	0.01	N/A	N/A	N/A	Adam	N/A
Almaslukh et al. [98]	0.0003	N/A	N/A	0.05	Adam	N/A
Lee et al. [99]	N/A	200	64	0.5	Adam	Cross entropy
Ye et al. [112]	N/A	50	N/A	N/A	RMSprop	N/A
Ye et al. [113]	0.0001	N/A	N/A	N/A	SGD	Cross entropy + Triplet loss
Chen et al. [114]	0.001	N/A	64	N/A	N/A	N/A
Alnujaim et al. [115]	0.0005	3000	64	0.4	SGD	Cross entropy
Erol et al. [116]	0.0002	3000		0.15	Adam	N/A
Wu and Ye [117]	0.0005	250	64	N/A	SGD	Cross entropy

*Note: N/A = Not Available.

smartphones for data collection and transfer purposes as modern smartphones contain almost all of the commonly used sensing modalities. The sensors were also positioned in various positions of the body. The datasets that used smartphones for data collection generally positioned the smartphone in front thigh positions.

Large-scale vision-based datasets that are collected from the video sharing site YouTube have numerous subjects. The system proposed in Ref. [128] contains 1 million videos containing various activities, while [129] contains 8 million videos of numerous subjects performing activities across a lot of classes. The position of the camera sensors in these datasets also varies widely. We also consider the datasets containing videos from YouTube as spontaneous in nature, as the videos are made by numerous individuals in various circumstances and as the videos are not staged in most cases.

5. Discussions, open issues, and future works

The research works reviewed in the previous section are the current state of the art in their respective sensing modalities. The detailed discussions and current open issues are described here.

5.1. Discussions of reviewed systems

Research works from four sensing modalities were discussed in this review: multimodal sensing devices-based, smartphone-based, vision-based, and radar-based. In a broad sense, both multimodal sensing devices based, and smartphone-based modalities are mostly body-worn or carried by the users. Vision-based and radar modality-based solutions are stationary or surveillance oriented. While smartphone-based solutions provide high-level utility for sensing and processing data alongside network and GPS capabilities, multimodal sensing solutions provide low-power always-on monitoring capabilities without requiring frequent recharging. Sometimes it becomes burdensome to carry the extra wearable devices for the users in the case of multimodal sensing devices. As smartphones require more processing power, for this why, daily-basis battery charging is required in such systems. Smartphone and multimodal solutions often monitor single users at a time, whereas radar and vision-based solutions can monitor multiple persons in an environment. The setup of radar-based systems is relatively harder compared to vision-based sensing modalities. The major problem of vision-based sensing modalities is that it is affected by environmental conditions such as the quality of background illumination. Fig. 5 provides a percentile representation of the modalities of the works reviewed in this paper.

Most of the systems described in this review used benchmark data for their experiments, only a few of the developed systems utilized their collected data from the participants, and almost all the systems used different datasets. The highest accuracy (greater than 99%) was found from the reviewed systems [57,115] for multi-sensing devices data (SHO [81]), and radar data (environmental data collected by Doppler sensor) respectively. In vision-based systems, the highest accuracy of 96.8% was achieved in Ref. [122] using the data from the RGB camera. However, comparing the systems is difficult due to the differences in the datasets. UCF101 [88] and HMDB51 [89] are the most used datasets in the reviewed vision-based systems but the obtained performances from the literature for UCF101 are comparatively better. The found evaluation metric (accuracy) for the frameworks presented in Refs. [72,125,132], and [133] are 94.94%, 88%, 91.5%, and 98.30%, respectively for UCF101 dataset. The highest accuracy is found from Ref. [72], and [133] using two-stream model fusion by SVM, as well as HOG, CNN-LSTM, and k-NN, respectively, for UCF101 dataset. The systems developed in Refs. [72,125], and [132] used HMDB51 dataset and achieved an accuracy of 70.6%, 59.4%, and 65.9% using CNN and SVM, two-stream model fusion by SVM, and 2-D CNN respectively. In general, vision-based human activity recognition works with huge datasets such as YouTube 1 M Sports [128] and YouTube 8 M [129]. The lowest F1 Score (68%) is obtained from Ref. [59] for multimodal real-time collected data using the fusion of CNN and RNN. The frequently used datasets are WISDM [90], and UCI [79] in this study. The WISDM dataset is used in Refs. [73,74,102–105] for human activity recognition. The frameworks presented in Ref. [74], and [104] mentioned F1 Score as only the performance measure and the values are 98.61%, and 95.69% using Reset and attention, as well as 2-D CNN and attention architectures respectively. The accuracy values of the systems developed in Refs. [57,67,105], and [106] are 97.59%, 93.41%, 97.05%, and 97.8% through the use of DCNN+, 2-D CNN and attention, CNN and BiLSTM, as well as encoder-decoder architecture, respectively. The performances of the developed frameworks for the same dataset are varied due to the use of the different deep learning architectures. A few of the HAR systems [98,100] provide real-time facilities i.e. ready for practical uses.

In this review, the human activity recognition systems are considered that mainly used CNNs for feature extraction or classification purposes. Some systems exploited only 2-D CNN for feature extraction and classification purposes, a few of the frameworks introduced fusion architecture such as the combination of CNN and other networks, and some systems employed attention mechanism along with CNN architecture for human activity recognition. The performances of each reviewed system varied as they utilized different benchmark datasets for experiments. The systems that are developed based on only CNN for feature extraction and classification purposes show comparatively less performance than the hybrid and attention-based frameworks. As a single network is used in such systems, it is not more generalized. The fusion architecture increased the generalizability as well as the performance by using different networks for feature extraction and classification. Moreover, the attention-based frameworks enhanced the performances of human activity recognition as the attention networks put more importance on the relevant features and remove the unnecessary (noisy) features at the time of feature extraction. Only the significant features are passed to the classification module; thus, the performances have been increased in the

case of attention-based frameworks. For example, we mentioned the performance (accuracy) of the widely used benchmark dataset called WISDM [90]. The accuracy value of the systems developed in Ref. [73] is 98.85% that used residual network and dual attention for recognition. The framework introduced in Ref. [105] used CNN and BiLSTM networks and obtained an accuracy of 98.53%. Besides, the temporal CNN employed in Ref. [101] found an accuracy of 98.2% from the same dataset.

Table 8 demonstrates the hyperparameters of the CNN structures that are used to develop each HAR system in this review. Some of the developed systems did not mention the hyperparameters in their work, and some of the frameworks mentioned a few parameters. Among the hyperparameters, the maximum times used batch size, dropout, optimizer, and loss are 64, 0.5, Adam, and cross-entropy, respectively.

5.2. Open issues

Although the existing human activity recognition-based techniques are excellent in recognizing atomic and basic activities in single-subject scenarios, they still struggle with HAR in various complex real-life scenarios. Some of the open challenges in HAR systems are presented below.

Complexity of modeling composite activities: While basic activities such as walking, running, sitting down are relatively easier to recognize, composite activities such as doing exercises containing multiple routines are significantly harder to model and recognize. There is a significant lack of datasets containing activity data on such concurrent activities.

Lack of activity data in multi-person scenarios: The majority of HAR datasets contain data on activities of a single person in experimental environments. Thus, the majority of the HAR systems are also capable of performing HAR on single-subject scenarios. However, the real world is filled in instances where multi-subject HAR is necessary such as in shops, kitchens, and living rooms, or in the case of multiple subjects involved in atomic activities such as handshaking, and hugging.

Lack of activity data in group scenarios: Most of the widely used HAR datasets represent activity data from singular humans performing various activities. Thus, most human activity recognition systems also detect the activity of singular humans. However, humans in real-life perform various activities in groups. There are no datasets that focus on activity data in group activities such as queues in shops, people walking, or jogging together.

Lack of contextual information in activity data: As human activity recognition is very closely related to human behavior understanding, context plays a huge role in human activity recognition. The same activity might be interpreted differently based on the context in which it is performed. Contextual information such as "where" (location) context or "when" (time) context can play huge roles in understanding human activity and behavior. For example, lying down in common resting places such as the sofa or beds can be interpreted as resting action, while lying down in bathrooms or kitchens can be interpreted as fall activity or signs of stroke. Similarly, a person watching tv or walking around the house after midnight might be interpreted as insomniac behavior. Another important contextual information is the repetitiveness of an activity. For example, activities such as eating too many times or too little a day can be interpreted as early symptoms of depression or mental instability. Human-human interaction or human-object interaction is also a great indicator of the meanings behind complex activities. Although contextual information is very important, we have not found datasets or human activity recognition systems that take contextual information into consideration.

Lack of relevant information in activity datasets: There are a lot of open datasets related to human activity recognition using various sensing modalities. However, the datasets are not standardized, meaning all of the datasets do not present similar levels of details on the test subjects, test environments, and proper data size. For example, most of the datasets did not provide very important information on the test subjects such as age range, height, and weight.

Lack of reproducibility of current works: In a lot of the reviewed works, very specific details on the network architectures and their hyperparameters are not provided. None of the papers provided the code behind the experimentations and model implementations. This makes these systems very hard to reproduce for future experimentation or benchmarking.

Lack of datasets containing concurrent activity data: The majority of the HAR systems also perform on the basis that a single human is taking part in a single activity at any given time. However, in real-life scenarios, subjects often perform multi-task and take part in concurrent activities, such as walking while drinking coffee, having snacks, or drinks while reading.

Lack of benchmark datasets containing data from radar sensing modality: While there is an abundance of benchmark datasets containing activity data from the various sensing modalities such as the accelerometer, gyroscope, RGB and RGB-D cameras, smartphones, industrial sensors, there is a clear lack of benchmark datasets containing activity data from radar sensing modalities.

Class imbalance for specific activities: While data for common tasks such as walking, talking, running, and swimming are very common, data for other abnormal activities such as accidental falls from various positions are very rare even in specialized datasets. Thus, in many cases, a class imbalance exists among the data for different types of activities.

Inter-activity variability: Variability is another open challenge for HAR systems. All humans do not perform a single task in the same way. As the CNN models generalize on the training dataset, in test cases, if subjects are performing the same tasks in different ways, then the models become unable to properly recognize them.

Inherent challenges of underlying technologies*:* The HAR systems based on radar signals have some shortcomings such as the lack of portability, very costly hardware requirements, and environmental interference. Similarly, smartphone-based and wearable sensors-based HAR systems have constraints on wearability. The HAR systems based on multiple sensing modalities are also prone to noise in the data, thus making the data harder to interpret and generalize too.

Computational complexity and requirements of CNN systems: CNN-based systems are also very computationally expensive for both training and inference purposes. While techniques exist to make CNNs more applicable for low-power devices such as embedded systems or edge devices, these methods often affect the overall performance of the systems.

Difference between real-world data and experimental settings: Most of the HAR systems were based on data collected in experimental environments. However, real-world activities are very complex and thus harder to model. In addition to composite activities, real-world environments also add complexities such as occlusion, interference, and noise. For example, the type of surface, clothing, and previous history of injuries or surgeries affect the activity of individuals in real life.

Lack of standardization: Various research works use various testing measures and benchmarks. Some HAR systems use parts of datasets, while others use different testing criteria. Thus, it becomes very hard to perform quantitative comparison among the systems and to perform a proper evaluation.

Data privacy or lack thereof: As HAR systems deal with very confidential real-time data of humans, maintaining the privacy of the collected data is another open challenge. Most of the networked HAR systems are prone to malicious attempts due to the lack of implementation of proper data privacy policies. However, the addition of complex data privacy policies, adding the latencies of the systems, are affecting the performance of real-time HAR systems.

Lack of explainability of CNN models: CNN models are incredibly complex mathematical models. CNN models were often compared to black boxes in the past due to the automatic feature extraction and

learning mechanism. It is very difficult to explain and visualize the extracted features from the hidden or intermediate deep layers of CNN models for human activity recognition. This lack of explainability results in lesser general adoption in sensitive use cases.

Complexity of training generalized CNN models: While CNN models can provide major performance improvements over other methods, they are susceptible to overfitting to specific training datasets. Overfitted models learn the data points of the training dataset instead of learning the latent representation. As a result, overfitted CNN models trained on specific datasets tend to severely underperform in the real world or unseen data use cases. Model generalization is harder to achieve than model specialization.

Complexity of deployment in edge computing and IoT scenarios: Recent research developments are focused on training and deploying complex deep learning models and architectures in resource-constrained edge devices connected via IoT networks. Edge devices are all around us. Mass deployment of HAR applications in edge devices would transform human lives in unprecedented ways. However, the challenges of such mass adoption include power and computational constraints of edge devices, performance loss while downsizing models through quantization, lack of available and accessible software support to design and deploy such systems, network load balancing and scalability issues, and lack of specialized network protocols or packet design for transferring these kinds of data.

5.3. Future works

HAR systems are continuously evolving along with their underlying technologies. The following research directions can be pursued to further advance the current works in this field.

Generative models for handling class imbalance: Recently, generative models such as GANs [141,142] are being widely used to generate photo-realistic fabricated data. These generative models can be tuned to generate data related to imbalanced classes. This might result in better models that would be capable of recognizing abnormal activities and thus saving lives.

Future activity prediction: Future activity prediction is an expansion of HAR that enables the prediction of probable activities by monitored humans. Future activity prediction has applications in law enforcement and driver behavior detection. As human activities are done sequentially in time, using other technologies (ex. Brain-computer interfaces, fMRI, EEG) and mechanisms (ex. attention mechanism) in conjunction with CNNs might result in effective future activity prediction systems.

Incorporating contextual information with activity data: Incorporating contextual information alongside sensor data in future open datasets would greatly aid complex human activity recognition systems. For example, incorporating timestamps, location, and audio with the relevant sensor data would aid in recognizing complex human actions. Similarly, audio and sentiment analysis systems can be used alongside human activity recognition systems to properly consider environmental contexts.

Standardization of representation of relevant data in open datasets: As stated earlier, a lot of the open datasets do not present relevant information such as the age range, height, weight, physical deficiencies, or known medical conditions of subjects in the manuscripts. However, different age groups perform similar activities very differently. Similarly, previous medical accidents or surgeries can significantly alter the ability of a subject to perform an action in a specific way. This information can play a huge role in designing criteria-specific or age-specific human activity recognition systems. A standardized data representation system for future human activity recognition datasets should be developed so that the relevant information can be easily accessed and incorporated in designing human activity recognition systems.

Creating robust human activity recognition systems: Environmental effects such as the choice of clothing and apparel, and type of surface greatly affect the activity of individuals. These changes are also represented in their relevant sensor data. However, the open datasets on human activity recognition are mostly collected in laboratory test environments that do not incorporate these differences. Thus, in the future, research efforts can be focused on developing real-world activity datasets on different environments and robust human activity recognition systems that are impervious to the changes in the environment.

Focusing on training explainable and generalized CNN models: Recent research advances have been focused on developing explainability methods for CNN models. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [143] and Shapley Additive Explanations (SHAP) [144] can be used in future works to explain the extracted features by the CNN models from various datasets. For training generalized CNN models, data partition methods such as k-fold cross validation, leave-one-out cross validation, hold-out portions of datasets as well as training methods such as regularization, dropout layers, and early stopping can be utilized. Steps should be taken to reduce data leaks between train and test sets to ensure the validity of performance metrics. Reproducibility should also be a major focus for future research.

Developing novel TinyML methods: As conventional CNN methods are computationally expensive to both train and deploy, future research should be focused on developing novel training and deployment methods for embedded systems using TinyML [145,146]. TinyML systems are both power-efficient and computationally efficient. Power efficiency is achieved through methods such as model pruning and weight clustering. Computational efficiency is achieved through weights and activation quantization and reduced operation compatibility.

5.4. Applications

Human activity recognition systems have widespread applications in various fields including smart healthcare systems [147,148], surveillance frameworks [149,150], and entertainment systems [151]. The main aim of human activity recognitions systems is to understand and analyze classified human actions and to interpret their semantic meaning in different domains. Human activities consist of simple atomic actions [152] such as walking, breathing; complex actions such as dancing, exercising; interpersonal interactions like handshaking, waving; or human-object interaction such as preparing meals, and working in production lines. As human activity recognition has very close ties with human behavior understanding and modeling, human activity recognition systems have application in diverse application domains. In this section, we briefly discuss some of the prevalent application domains of human activity recognition.

5.4.1. Healthcare Systems

Human activity recognition systems are widely used in healthcare systems to analyze and interpret patient activities for facilitating healthcare and essential workers to monitor, diagnose, and care for patients [153]. This results in improved accuracy of diagnosis and case, the decreased workload for healthcare staff, increased quality of service received by patients, decreased hospital stays, decreased medical cost, and decreased chances of serious injury. Human activity recognition systems are used in various medical use cases such as automatic fall detection [154,155], and response systems [156,157] for detecting accidental falls and providing immediate response services; respiratory behavior modeling systems to recognize and diagnose sleep disorders [158], cardiovascular diseases [159], and stroke [160]; medication intake monitoring systems to ensure proper usage of medicine [161,162]; hand movement monitoring system to recognize and diagnose eating disorders [163,164]; exercise-aid systems to guide in proper postures during regular exercises [165]; hand gesture recognition system for sign language-based interaction [166] and automatic wheelchair movement [167].

5.4.2. Surveillance Systems

Surveillance systems are another application domain that extensively utilizes human activity recognition systems. Activity recognition systems are used in surveillance scenarios to track and monitor individuals and crowds, thus supporting security personnel to observe and detect suspicious activities and threats. Human activity recognition systems have different use cases in surveillance systems such as gait based long-range person recognition [168,169] and authentication [170] to detect and recognize specific individuals from a long distance based on gait patterns; driver drowsiness detection systems to ensure proper driver behavior [171] and to reduce road accidents due to driver inattention [172]; automatic drowning detection systems [173] in swimming pools to save lives of swimmers and to reduce chances of long-term damage; loitering detection systems to detect suspicious loitering behavior or erratic movements of individuals around important public spaces [174]; suspicious activity detection systems to detect violent interpersonal behaviors [175].

5.4.3. Entertainment Systems

Human activity recognition systems are widely used in entertainment systems to both monitor and aid referees and players in sports and to interact with computer games in fun ways. Some of the use cases of activity recognition systems in entertainment systems are as follows: accurate automatic timer systems that detect the start and end time of an activity such as swimming [176], diving [177]; pose-estimation systems for detecting and scoring real-life dance moves [178,179], navigating in 3-D spaces [180], interacting in virtual environments [181]; movement recognition system for detecting various types of strokes and events during tennis games [182,183].

6. Conclusion

Human activity recognition systems are essential tools for humanity as they enable the recording of general human activities through different sensing modalities and the monitoring, analysis, and assistance of daily life through capable computing systems. Human activity recognition systems have numerous applications in various important fields such as healthcare systems, surveillance systems, and entertainment systems. This review work is an exploration of the use of convolutional neural networks in human activity recognition systems through the presented sensing modalities. The major four sensing modalities: multi-sensor-based systems, smartphone-based systems, radar-based systems, and vision-based systems are demonstrated in this review. The different and effective use of various CNN architectures and techniques such as 1-D CNNs, inception blocks, dense blocks, two-stream convolutional networks, 3-D CNNs, specialized features such as trajectory pooled deep convolutional detector are highlighted in this review. The reviewed systems are presented in the light of the issues they solve, their strengths, their weaknesses, and their performance. Available hyperparameters of the reviewed systems are also presented. We also presented brief details on the available public datasets containing data collected through various sensing modalities that are frequently used in the reviewed systems and the field in general. Finally, we discuss the open challenges, the applications as well as some potential solutions.

References

[1] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, G. Fortino, Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges, Inf. Fusion 80 (2022) 241–265, https://doi.org/10.1016/j.inffus.2021.11.006.

[2] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, Y. Liu, Deep learning for sensor-based human activity recognition, ACM Comput. Surv. 54 (2021) 1–40, https://doi.org/10.1145/3447744.

[3] S. Jiang, P. Kang, X. Song, B. Lo, P.B. Shull, Emerging wearable interfaces and algorithms for hand gesture recognition: a survey, IEEE Rev. Biomed. Eng. 15 (2022) 85–102, https://doi.org/10.1109/RBME.2021.3078190.

[4] P. Das Sakshi, S. Jain, C. Sharma, V. Kukreja, Deep learning: an application perspective, in: Lect. Notes Networks Syst., 2022, pp. 323–333, https://doi.org/10.1007/978-981-16-4284-5_28.

[5] Y. Liu, Q. Zhang, W. Chen, Massive-scale complicated human action recognition: theory and applications, Future Generat. Comput. Syst. 125 (2021) 806–811, https://doi.org/10.1016/j.future.2021.06.060.

[6] F. Luo, S. Poslad, E. Bodanese, Kitchen activity detection for healthcare using a low-power radar-enabled sensor network, in: ICC 2019 - 2019 IEEE Int. Conf. Commun., IEEE, 2019, pp. 1–7, https://doi.org/10.1109/ICC.2019.8761484.

[7] E. Alam, A. Sufian, P. Dutta, M. Leo, Vision-based human fall detection systems using deep learning: a review, Comput. Biol. Med. 146 (2022), 105626, https://doi.org/10.1016/j.compbiomed.2022.105626.

[8] A. Sanchez-Comas, K. Synnes, J. Hallberg, Hardware for recognition of human activities: a review of smart home and AAL related technologies, Sensors 20 (2020) 4227, https://doi.org/10.3390/s20154227.

[9] M. Rawashdeh, M.G. Al Zamil, S. Samarah, M.S. Hossain, G. Muhammad, A knowledge-driven approach for activity recognition in smart homes based on activity profiling, Future Generat. Comput. Syst. 107 (2020) 924–941, https://doi.org/10.1016/j.future.2017.10.031.

[10] L. Schrader, A. Vargas Toro, S. Konietzny, S. Rüping, B. Schäpers, M. Steinböck, C. Krewer, F. Müller, J. Güttler, T. Bock, Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people, J. Popul. Ageing 13 (2020) 139–165, https://doi.org/10.1007/s12062-020-09260-z.

[11] M. Jacob Rodrigues, O. Postolache, F. Cercas, Physiological and behavior monitoring systems for smart healthcare environments: a review, Sensors 20 (2020) 2186, https://doi.org/10.3390/s20082186.

[12] W. Zhang, C. Su, C. He, Rehabilitation exercise recognition and evaluation based on smart sensors with deep learning framework, IEEE Access 8 (2020), https://doi.org/10.1109/ACCESS.2020.2989128, 77561–77571.

[13] A. Ullah, K. Muhammad, I.U. Haq, S.W. Baik, Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments, Future Generat. Comput. Syst. 96 (2019) 386–397, https://doi.org/10.1016/j.future.2019.01.029.

[14] L. Martínez-Villaseñor, H. Ponce, A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction, Int. J. Distributed Sens. Netw. 15 (2019), 155014771985398, https://doi.org/10.1177/1550147719853987.

[15] A.S. M, N. Thillaiarasu, A survey on different computer vision based human activity recognition for surveillance applications, in: 2022 6th Int. Conf. Comput. Methodol. Commun., IEEE, 2022, pp. 1372–1376, https://doi.org/10.1109/ICCMC53470.2022.9753931.

[16] W. Zheng, L. Yan, C. Gou, F.-Y. Wang, Meta-learning meets the Internet of Things: graph prototypical models for sensor-based human activity recognition, Inf. Fusion 80 (2022) 1–22, https://doi.org/10.1016/j.inffus.2021.10.009.

[17] P. Pareek, A. Thakkar, A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. 54 (2021) 2259–2322, https://doi.org/10.1007/s10462-020-09904-8.

[18] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recogn. 47 (2014) 3343–3361, https://doi.org/10.1016/j.patcog.2014.04.018.

[19] D. Riboni, M. Murtas, Sensor-based activity recognition: one picture is worth a thousand words, Future Generat. Comput. Syst. 101 (2019) 709–722, https://doi.org/10.1016/j.future.2019.07.020.

[20] N. Rashid, B.U. Demirel, M.A. Al Faruque, AHAR, AHAR, Adaptive CNN for energy-efficient human activity recognition in low-power edge devices, IEEE Internet Things J. 9 (2022) 13041–13051, https://doi.org/10.1109/JIOT.2022.3140465.

[21] E. Fridriksdottir, A.G. Bonomi, Accelerometer-based human activity recognition for patient monitoring using a deep neural network, Sensors 20 (2020) 6424, https://doi.org/10.3390/s20226424.

[22] H. Arab, I. Ghaffari, L. Chioukh, S.O. Tatu, S. Dufour, A convolutional neural network for human motion recognition and classification using a millimeter-wave Doppler radar, IEEE Sensor. J. 22 (2022) 4494–4502, https://doi.org/10.1109/JSEN.2022.3140787.

[23] X. Fan, F. Wang, F. Wang, W. Gong, J. Liu, When RFID meets deep learning: exploring cognitive intelligence for activity identification, IEEE Wireless Commun. 26 (2019) 19–25, https://doi.org/10.1109/MWC.2019.1800405.

[24] Y. Wang, S. Cang, H. Yu, A survey on wearable sensor modality centred human activity recognition in health care, Expert Syst. Appl. 137 (2019) 167–190, https://doi.org/10.1016/j.eswa.2019.04.057.

[25] T. Özyer, D.S. Ak, R. Alhajj, Human action recognition approaches with video datasets—a survey, Knowl. Base Syst. 222 (2021), 106995, https://doi.org/10.1016/j.knosys.2021.106995.

[26] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods, Sensors 19 (2019) 1005, https://doi.org/10.3390/s19051005.

[27] A.M. F, S. Singh, Computer vision-based survey on human activity recognition system, challenges and applications, in: 2021 3rd Int. Conf. Signal Process. Commun., IEEE, 2021, pp. 110–114, https://doi.org/10.1109/ICSPC51351.2021.9451736.

[28] V. Sharma, M. Gupta, A.K. Pandey, D. Mishra, A. Kumar, A review of deep learning-based human activity recognition on benchmark video datasets, Appl. Artif. Intell. 36 (2022) 2093705, https://doi.org/10.1080/08839514.2022.2093705.

[29] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, N. Alshurafa, Deep learning in human activity recognition with wearable sensors: a review on advances, Sensors 22 (2022) 1476, https://doi.org/10.3390/s22041476.
[30] I.M. Pires, F. Hussain, G. Marques, N.M. Garcia, Comparison of machine learning techniques for the identification of human activities from inertial sensors available in a mobile device after the application of data imputation techniques, Comput. Biol. Med. 135 (2021), 104638, https://doi.org/10.1016/j.compbiomed.2021.104638.
[31] A. Maurya, R.K. Yadav, M. Kumar, Saumya, Comparative Study of Human Activity Recognition on Sensory Data Using Machine Learning and Deep Learning, 2021, pp. 63–71, https://doi.org/10.1007/978-981-33-6307-6_8.
[32] F. Alshehri, G. Muhammad, A comprehensive survey of the Internet of things (IoT) and AI-based smart healthcare, IEEE Access 9 (2021) 3660–3678, https://doi.org/10.1109/ACCESS.2020.3047960.
[33] D.V. Medhane, A.K. Sangaiah, M.S. Hossain, G. Muhammad, J. Wang, Blockchain-enabled distributed security framework for next-generation IoT: an edge cloud and software-defined network-integrated approach, IEEE Internet Things J. 7 (2020) 6143–6149, https://doi.org/10.1109/JIOT.2020.2977196.
[34] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: a survey, Pattern Recogn. Lett. 119 (2019) 3–11, https://doi.org/10.1016/j.patrec.2018.02.010.
[35] E. Ramanujam, T. Perumal, S. Padmavathi, Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review, IEEE Sensor. J. 21 (2021) 13029–13040, https://doi.org/10.1109/JSEN.2021.3069927.
[36] J.K. Dhillon, Chandni, A.K.S. Kushwaha, A recent survey for human activity recognition based on deep learning approach, in: 2017 Fourth Int. Conf. Image Inf. Process, IEEE, 2017, pp. 1–6, https://doi.org/10.1109/ICIIP.2017.8313715.
[37] S. Ramasamy Ramamurthy, N. Roy, Recent trends in machine learning for human activity recognition—a survey, WIREs Data Min. Knowl. Discov. 8 (2018) e1254, https://doi.org/10.1002/widm.1254.
[38] K.K. Verma, B.M. Singh, A. Dixit, A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system, Int. J. Inf. Technol. 14 (2022) 397–410, https://doi.org/10.1007/s41870-019-00364-0.
[39] A. Biswal, S. Nanda, C.R. Panigrahi, S.K. Cowlessur, B. Pati, Human activity recognition using machine learning: a review, Adv. Intell. Syst. Comput. (2021) 323–333, https://doi.org/10.1007/978-981-33-4299-6_27.
[40] P. Kumari, L. Mathew, P. Syal, Increasing trend of wearables and multimodal interface for human activity monitoring: a review, Biosens. Bioelectron. 90 (2017) 298–307, https://doi.org/10.1016/j.bios.2016.12.001.
[41] U. Alrazzak, B. Alhalabi, A Survey on Human Activity Recognition Using Accelerometer Sensor, in: 2019 Jt. 8th Int. Conf. Informatics, Electron. Vis. 2019 3rd Int. Conf. Imaging, Vis. Pattern Recognit., IEEE, 2019, pp. 152–159, https://doi.org/10.1109/ICIEV.2019.8858578.
[42] M. Cornacchia, K. Ozcan, Y. Zheng, S. Velipasalar, A survey on activity detection and classification using wearable sensors, IEEE Sensor. J. 17 (2017) 386–403, https://doi.org/10.1109/JSEN.2016.2628346.
[43] D.R. Beddiar, B. Nini, M. Sabokrou, A. Hadid, Vision-based human activity recognition: a survey, Multimed. Tool. Appl. 79 (2020) 30509–30555, https://doi.org/10.1007/s11042-020-09004-3.
[44] A. Bux, P. Angelov, Z. Habib, Vision based human activity recognition: a review, Adv. Intell. Syst. Comput. (2017) 341–371, https://doi.org/10.1007/978-3-319-46562-3_23.
[45] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71, https://doi.org/10.1136/bmj.n71.
[46] M.U. Khan, S. Sherin, M.Z. Iqbal, R. Zahid, Landscaping systematic mapping studies in software engineering: a tertiary study, J. Syst. Software 149 (2019) 396–436, https://doi.org/10.1016/j.jss.2018.12.018.
[47] A. Gumaei, M.M. Hassan, A. Alelaiwi, H. Alsalman, A hybrid deep learning model for human activity recognition using multimodal body sensing data, IEEE Access 7 (2019) 99152–99160, https://doi.org/10.1109/ACCESS.2019.2927134.
[48] T. Mahmud, A.Q.M.S. Sayyed, S.A. Fattah, S.-Y. Kung, A novel multi-stage training approach for human activity recognition from multimodal wearable sensor data using deep neural network, IEEE Sensor. J. 21 (2021) 1715–1726, https://doi.org/10.1109/JSEN.2020.3015781.
[49] R. Abdel-Salam, R. Mostafa, M. Hadhood, Human activity recognition using wearable sensors: review, challenges, evaluation benchmark, Commun. Comput. Inf. Sci. 1370 (2021) 1–15, https://doi.org/10.1007/978-981-16-0575-8_1.
[50] F. Serpush, M.B. Menhaj, B. Masoumi, B. Karasfi, Wearable sensor-based human activity recognition in the smart healthcare system, Comput. Intell. Neurosci. 2022 (2022) 1–31, https://doi.org/10.1155/2022/1391906.
[51] A. Mimouna, A. Ben Khalifa, A survey of human action recognition using accelerometer data, in: Smart Sensors, Meas. Instrum., 2021, pp. 1–32, https://doi.org/10.1007/978-3-030-71225-9_1.
[52] Khimraj, P.K. Shukla, A. Vijayvargiya, R. Kumar, Human activity recognition using accelerometer and gyroscope data from smartphones, in: 2020 Int. Conf. Emerg. Trends Commun. Control Comput, IEEE, 2020, https://doi.org/10.1109/ICONC345789.2020.9117456, 1–6.
[53] A.K.M. Masum, E.H. Bahadur, A. Shan-A-Alahi, M.A. Uz Zaman Chowdhury, M. R. Uddin, A. Al Noman, Human activity recognition using accelerometer, gyroscope and magnetometer sensors: deep neural network approaches, in: 2019 10th Int. Conf. Comput. Commun. Netw. Technol, IEEE, 2019, pp. 1–6, https://doi.org/10.1109/ICCCNT45670.2019.8944512.
[54] M.S. Afzali Arani, D.E. Costa, E. Shihab, Human activity recognition: a comparative study to assess the contribution level of accelerometer, ECG, and PPG signals, Sensors 21 (2021) 6997, https://doi.org/10.3390/s21216997.
[55] K. Nurhanim, I. Elamvazuthi, L.I. Izhar, G. Capi, S. Su, EMG signals classification on human activity recognition using machine learning algorithm, in: 2021 8th NAFOSTED Conf. Inf. Comput. Sci., IEEE, 2021, pp. 369–373, https://doi.org/10.1109/NICS54270.2021.9701461.
[56] S. Ha, J.-M. Yun, S. Choi, Multi-modal Convolutional Neural Networks for Activity Recognition, in: 2015 IEEE Int. Conf. Syst. Man, Cybern, IEEE, 2015, pp. 3017–3022, https://doi.org/10.1109/SMC.2015.525.
[57] W. Jiang, Z. Yin, Human activity recognition using wearable sensors by deep convolutional neural networks, in: Proc. 23rd ACM Int. Conf. Multimed. - MM '15, ACM Press, New York, New York, USA, 2015, pp. 1307–1310, https://doi.org/10.1145/2733373.2806333.
[58] J.-H. Choi, J.-S. Lee, EmbraceNet: A robust deep learning architecture for multimodal classification, Inf. Fusion 51 (2019) 259–270, https://doi.org/10.1016/j.inffus.2019.02.010.
[59] J.-H. Choi, J.-S. Lee, Confidence-based deep multimodal fusion for activity recognition, in: Proc. 2018 ACM Int. Jt. Conf. 2018 Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput. - UbiComp '18, ACM Press, New York, New York, USA, 2018, pp. 1548–1556, https://doi.org/10.1145/3267305.3267522.
[60] C.-T. Yen, J.-X. Liao, Y.-K. Huang, Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms, IEEE Access 8 (2020) 174105–174114, https://doi.org/10.1109/ACCESS.2020.3025938.
[61] I.A. Lawal, S. Bano, Deep human activity recognition with localisation of wearable sensors, IEEE Access 8 (2020) 155060–155070, https://doi.org/10.1109/ACCESS.2020.3017681.
[62] S. Ha, S. Choi, Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors, in: 2016 Int. Jt. Conf. Neural Networks, IEEE, 2016, pp. 381–388, https://doi.org/10.1109/IJCNN.2016.7727224.
[63] J.B. Yang, M.N. Nguyen, P.P. San, X.L. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, IJCAI Int. Jt. Conf. Artif. Intell. (2015) 3995–4001.
[64] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62, https://doi.org/10.1016/j.neucom.2021.03.091.
[65] K. Muhammad, Mustaqeem, A. Ullah, A.S. Imran, M. Sajjad, M.S. Kiran, G. Sannino, V.H.C. de Albuquerque, Human action recognition using attention based LSTM network with dilated CNN features, Future Generat. Comput. Syst. 125 (2021) 820–830, https://doi.org/10.1016/j.future.2021.06.045.
[66] R.A. Hamad, M. Kimura, L. Yang, W.L. Woo, B. Wei, Dilated causal convolution with multi-head self attention for sensor human activity recognition, Neural Comput. Appl. 33 (2021) 13705–13722, https://doi.org/10.1007/s00521-021-06007-5.
[67] K. Wang, J. He, L. Zhang, Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors, IEEE Sensor. J. 19 (2019) 7598–7604, https://doi.org/10.1109/JSEN.2019.2917225.
[68] T.-H. Tan, C.-J. Huang, M. Gochoo, Y.-F. Chen, Activity recognition based on FR-CNN and attention-based LSTM network, in: 2021 30th Wirel. Opt. Commun. Conf., IEEE, 2021, pp. 146–149, https://doi.org/10.1109/WOCC53213.2021.9603203.
[69] Y. Liu, H. Zhang, D. Xu, K. He, Graph transformer network with temporal kernel attention for skeleton-based action recognition, Knowledge-Based Syst. 240 (2022) 108146, https://doi.org/10.1016/j.knosys.2022.108146.
[70] H. Wang, Deeply-learned and spatial–temporal feature engineering for human action understanding, Futur, Gener. Comput. Syst. 123 (2021) 257–262, https://doi.org/10.1016/j.future.2021.04.021.
[71] Y.A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M.I. Oros-Flores, D.-L. Almanza-Ojeda, Human activity recognition using temporal convolutional neural network architecture, Expert Syst. Appl. 191 (2022), 116287, https://doi.org/10.1016/j.eswa.2021.116287.
[72] S. Zhu, Z. Fang, Y. Wang, J. Yu, J. Du, Multimodal activity recognition with local block CNN and attention-based spatial weighted CNN, J. Vis. Commun. Image Represent. 60 (2019) 38–43, https://doi.org/10.1016/j.jvcir.2018.12.026.
[73] W. Gao, L. Zhang, Q. Teng, J. He, H. Wu, DanHAR: dual attention network for multimodal human activity recognition using wearable sensors, Appl. Soft Comput. 111 (2020), 107728, https://doi.org/10.1016/j.asoc.2021.107728.
[74] Y. Tang, L. Zhang, Q. Teng, F. Min, A. Song, Triple cross-domain attention on human activity recognition using wearable sensors, IEEE Trans. Emerg. Top. Comput. Intell. (2022) 1–10, https://doi.org/10.1109/TETCI.2021.3136642.
[75] O. Banos, R. Garcia, J.A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, C. Villalonga, mHealthDroid: a novel framework for agile development of mobile health applications, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014, pp. 91–98, https://doi.org/10.1007/978-3-319-13105-4_14.
[76] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, G. Tröster, Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection, in: Wirel. Sens. Networks, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 17–33, https://doi.org/10.1007/978-3-540-77690-1_2.

[77] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, J. del R. Millan, Collecting complex activity datasets in highly rich networked sensor environments, in: 2010 Seventh Int. Conf. Networked, Sens. Syst., IEEE, 2010, pp. 233–240, https://doi.org/10.1109/INSS.2010.5573462.

[78] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, ACM Comput. Surv. 46 (2014) 1–33, https://doi.org/10.1145/2499621.

[79] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, A public domain dataset for human activity recognition using smartphones, in: ESANN 2013 Proceedings, 21st Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn., 2013, pp. 437–442.

[80] M. Zhang, A.A. Sawchuk, USC-HAD, A daily activity dataset for ubiquitous activity recognition using wearable sensors, in: UbiComp'12 - Proc. 2012 ACM Conf, Ubiquitous Comput., 2012, pp. 1036–1043.

[81] M. Shoaib, S. Bosch, O. Incel, H. Scholten, P. Havinga, Fusion of smartphone motion sensors for physical activity recognition, Sensors 14 (2014) 10146–10176, https://doi.org/10.3390/s140610146.

[82] A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov, R. Huerta, On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines, Sensor. Actuator. B Chem. 185 (2013) 462–477, https://doi.org/10.1016/j.snb.2013.05.027.

[83] T. Sztyler, H. Stuckenschmidt, W. Petrich, Position-aware activity recognition with wearable devices, Pervasive Mob. Comput. 38 (2017) 281–295, https://doi.org/10.1016/j.pmcj.2017.01.008.

[84] L. O, X.P.J. Reyes-Ortiz, D. Anguita, A. Ghio, UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set, 2012.

[85] S. Gaglio, G. Lo Re, M. Morana, Human activity recognition process using 3-D posture data, IEEE Trans. Human Mach Syst. 45 (2015) 586–597, https://doi.org/10.1109/THMS.2014.2377111.

[86] Jaeyong Sung, C. Ponce, B. Selman, A. Saxena, Unstructured Human Activity Detection from RGBD Images, in: 2012 IEEE Int. Conf. Robot. Autom, IEEE, 2012, pp. 842–849, https://doi.org/10.1109/ICRA.2012.6224591.

[87] Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conf. Comput. Vis. Pattern Recognit, IEEE, 2012, pp. 1290–1297, https://doi.org/10.1109/CVPR.2012.6247813.

[88] K. Soomro, A.R. Zamir, M. Shah, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, arXiv, 2012. CoRR abs/1212.0402.

[89] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, Proc. IEEE Int. Conf. Comput. Vis. (2011) 2556–2563, https://doi.org/10.1109/ICCV.2011.6126543.

[90] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Activity recognition using cell phone accelerometers, ACM SIGKDD Explor. Newsl. 12 (2011) 74–82, https://doi.org/10.1145/1964897.1964918.

[91] F.J. Ordóñez, P. de Toledo, A. Sanchis, Activity recognition using hybrid generative/discriminative models on home environments using binary sensors, Sensors 13 (2013) 5460–5477, https://doi.org/10.3390/s130505460.

[92] T.L.M. van Kasteren, G. Englebienne, B.J.A. Kröse, Human Activity Recognition from Wireless Sensor Network Data, Benchmark and Software, 2011, pp. 165–186, https://doi.org/10.2991/978-94-91216-05-3_8.

[93] D. Micucci, M. Mobilio, P. Napoletano, S.H.A.R. UniMiB, A dataset for human activity recognition using acceleration data from smartphones, Appl. Sci. 7 (2017) 1101, https://doi.org/10.3390/app7101101.

[94] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, in: 2012 16th Int. Symp. Wearable Comput, IEEE, 2012, pp. 108–109, https://doi.org/10.1109/ISWC.2012.13.

[95] D. Thakur, S. Biswas, Smartphone based human activity monitoring and recognition using ML and DL: a comprehensive survey, J. Ambient Intell. Hum. Comput. 11 (2020) 5433–5444, https://doi.org/10.1007/s12652-020-01899-y.

[96] G. Yuan, Z. Wang, F. Meng, Q. Yan, S. Xia, An overview of human activity recognition based on smartphone, Sens. Rev. 39 (2019) 288–306, https://doi.org/10.1108/SR-11-2017-0245.

[97] M. Straczkiewicz, P. James, J.-P. Onnela, A systematic review of smartphone-based human activity recognition methods for health research, Npj Digit. Med. 4 (2021) 148, https://doi.org/10.1038/s41746-021-00514-4.

[98] B. Almaslukh, A. Artoli, J. Al-Muhtadi, A robust deep learning approach for position-independent smartphone-based human activity recognition, Sensors 18 (2018) 3726, https://doi.org/10.3390/s18113726.

[99] Song-Mi Lee, Sang Min Yoon, Heeryon Cho, Human activity recognition from accelerometer data using Convolutional Neural Network, in: 2017 IEEE Int. Conf. Big Data Smart Comput, IEEE, 2017, pp. 131–134, https://doi.org/10.1109/BIGCOMP.2017.7881728.

[100] D. Ravi, C. Wong, B. Lo, G.-Z. Yang, A deep learning approach to on-node sensor data analytics for mobile or wearable devices, IEEE J. Biomed. Health Inf. 21 (2017) 56–64, https://doi.org/10.1109/JBHI.2016.2633287.

[101] D. Ravi, C. Wong, B. Lo, G.-Z. Yang, Deep learning for human activity recognition: a resource efficient implementation on low-power devices, in: 2016 IEEE 13th Int. Conf. Wearable Implant. Body Sens. Networks, IEEE, 2016, pp. 71–76, https://doi.org/10.1109/BSN.2016.7516235.

[102] Z.N. Khan, J. Ahmad, Attention induced multi-head convolutional neural network for human activity recognition, Appl. Soft Comput. 110 (2021), 107671, https://doi.org/10.1016/j.asoc.2021.107671.

[103] H. Zhang, Z. Xiao, J. Wang, F. Li, E. Szczerbicki, A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention, IEEE Internet Things J. 7 (2020) 1072–1080, https://doi.org/10.1109/JIOT.2019.2949715.

[104] G. Zheng, A novel attention-based convolution neural network for human activity recognition, IEEE Sensor. J. 21 (2021) 27015–27025, https://doi.org/10.1109/JSEN.2021.3122258.

[105] O. Nafea, W. Abdul, G. Muhammad, M. Alsulaiman, Sensor-based human activity recognition with spatio-temporal deep learning, Sensors 21 (2021) 2141, https://doi.org/10.3390/s21062141.

[106] N. Nair, C. Thomas, D.B. Jayagopi, Human activity recognition using temporal convolutional network, in: Proc. 5th Int. Work. Sensor-Based Act. Recognit. Interact., ACM, New York, NY, USA, 2018, pp. 1–8, https://doi.org/10.1145/3266157.3266221.

[107] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J.M. Hausdorff, N. Giladi, G. Troster, Wearable assistant for Parkinson's disease patients with the freezing of gait symptom, IEEE Trans. Inf. Technol. Biomed. 14 (2010) 436–446, https://doi.org/10.1109/TITB.2009.2036165.

[108] T. Sztyler, H. Stuckenschmidt, On-body localization of wearable devices: an investigation of position-aware activity recognition, in: 2016 IEEE Int. Conf. Pervasive Comput. Commun, IEEE, 2016, pp. 1–9, https://doi.org/10.1109/PERCOM.2016.7456521.

[109] J.W. Lockhart, G.M. Weiss, J.C. Xue, S.T. Gallagher, A.B. Grosner, T.T. Pulickal, Design considerations for the WISDM smart phone-based sensor mining architecture, in: Proc. Fifth Int. Work. Knowl. Discov. From Sens. Data - SensorKDD '11, ACM Press, New York, New York, USA, 2011, pp. 25–33, https://doi.org/10.1145/2003653.2003656.

[110] X. Li, Y. He, X. Jing, A survey of deep learning-based human activity recognition in radar, Rem. Sens. 11 (2019) 1068, https://doi.org/10.3390/rs11091068.

[111] A. Hanif, M. Muaz, A. Hasan, M. Adeel, Micro-Doppler based target recognition with radars: a review, IEEE Sensor. J. 22 (2022) 2948–2961, https://doi.org/10.1109/JSEN.2022.3141213.

[112] W. Ye, H. Chen, B. Li, Using an end-to-end convolutional network on radar signal for human activity classification, IEEE Sensor. J. 19 (2019) 12244–12252, https://doi.org/10.1109/JSEN.2019.2938997.

[113] W. Ye, H. Chen, Human activity classification based on micro-Doppler signatures by multiscale and multitask fourier convolutional neural network, IEEE Sensor. J. 20 (2020) 5473–5479, https://doi.org/10.1109/JSEN.2020.2971626.

[114] H. Chen, W. Ye, Classification of human activity based on radar signal using 1-D convolutional neural network, Geosci. Rem. Sens. Lett. IEEE 17 (2020) 1178–1182, https://doi.org/10.1109/LGRS.2019.2942097.

[115] I. Alnujaim, D. Oh, Y. Kim, Generative adversarial networks for classification of micro-Doppler signatures of human activity, Geosci. Rem. Sens. Lett. IEEE 17 (2020) 396–400, https://doi.org/10.1109/LGRS.2019.2919770.

[116] B. Erol, S.Z. Gurbuz, M.G. Amin, GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition, in: 2019 IEEE Radar Conf, IEEE, 2019, pp. 1–5, https://doi.org/10.1109/RADAR.2019.8835589.

[117] C. Wu, W. Ye, Generative adversarial network for radar-based human activities classification with low training data support, in: 2021 IEEE 4th Int. Conf. Electron. Inf. Commun. Technol., IEEE, 2021, pp. 415–419, https://doi.org/10.1109/ICEICT53123.2021.9531147.

[118] I. Alnujaim, S.S. Ram, D. Oh, Y. Kim, Synthesis of micro-Doppler signatures of human activities from different aspect angles using generative adversarial networks, IEEE Access 9 (2021) 46422–46429, https://doi.org/10.1109/ACCESS.2021.3068075.

[119] S. Sundar Ram, H. Ling, Simulation of human microDopplers using computer animation data, in: 2008 IEEE Radar Conf., IEEE, 2008, pp. 1–6, https://doi.org/10.1109/RADAR.2008.4720816.

[120] L.-F. Wu, Q. Wang, M. Jian, Y. Qiao, B.-X. Zhao, A comprehensive review of group Activity recognition in videos, Int. J. Autom. Comput. 18 (2021) 334–350, https://doi.org/10.1007/s11633-020-1258-8.

[121] T. Singh, D.K. Vishwakarma, Human activity recognition in video benchmarks: a survey, in: Lect. Notes Electr. Eng., 2019, pp. 247–259, https://doi.org/10.1007/978-981-13-2553-3_24.

[122] M.A. Gul, M.H. Yousaf, S. Nawaz, Z.U. Rehman, H. Kim, Patient monitoring by abnormal human activity recognition based on CNN architecture, Electronics 9 (12) 1993, doi:10.3390/electronics9121993.

[123] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis, . Pattern Recognit., . Pattern Recognit., 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[124] S. Shinde, A. Kothari, V. Gupta, ScienceDirect YOLO based human action recognition and localization, Procedia Comput. Sci. 133 (2019) 831–838, https://doi.org/10.1016/j.procs.2018.07.112.

[125] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. (2014) 1–9.

[126] D.A. Pisner, D.M. Schnyer, Support vector machine, in: Mach. Learn., Elsevier, 2020, pp. 101–121, https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

[127] D. Puchala, K. Stokfiszewski, M. Yatsymirskyy, Image statistics preserving encrypt-then-compress scheme dedicated for JPEG compression standard, Entropy 23 (2021) 421, https://doi.org/10.3390/e23040421.

[128] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.F. Li, Large-scale video classification with convolutional neural networks, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2014) 1725–1732, https://doi.org/10.1109/CVPR.2014.223.

[129] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, et al., YouTube-8M: a large-scale video classification benchmark, 2016 arXiv: 1609.08675.

[130] S. Ji, W. Xu, M. Yang, K. Yu, 3D Convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 221–231, https://doi.org/10.1109/TPAMI.2012.59.
[131] H. Wang, A. Kläser, C. Schmid, C.L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (2013) 60–79, https://doi.org/10.1007/s11263-012-0594-8.
[132] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis, . Pattern Recognit., 2015, pp. 4305–4314.
[133] F. Serpush, M. Rezaei, Complex human action recognition using a hierarchical feature reduction and deep learning-based method, SN Comput. Sci. 2 (2021) 94, https://doi.org/10.1007/s42979-021-00484-0.
[134] J. Basavaiah, C.M. Patil, Human activity detection and action recognition in videos using convolutional neural networks, J. Inf. Commun. Technol. 19 (2020) 157–183, https://doi.org/10.32890/jict2020.19.2.1.
[135] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.E. Bichot, C. Garcia, B. Sankur, Evaluation of video activity localizations integrating quality and quantity measurements, Comput. Vis. Image Understand. 127 (2014) 14–30, https://doi.org/10.1016/j.cviu.2014.06.014.
[136] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. - Int. Conf. Pattern Recognit., IEEE, 2004, pp. 32–36, https://doi.org/10.1109/ICPR.2004.1334462.
[137] M. Yang, S. Ji, W. Xu, J. Wang, F. Lv, K. Yu, Y. Gong, M. Dikmen, D.J. Lin, T. S. Huang, Detecting human actions in surveillance videos, in: 2009 TREC Video Retr, Eval. Noteb. Pap., 2009.
[138] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 2247–2253, https://doi.org/10.1109/TPAMI.2007.70711.
[139] A. Shahroudy, J. Liu, T.T. Ng, G. Wang, NTU RGB+D: a large scale dataset for 3D human activity analysis, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[140] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding, 2017 arXiv:1703.07475.
[141] A.Z.M. Faridee, A. Chakma, A. Misra, N. Roy, STranGAN, Adversarially-learnt Spatial Transformer for scalable human activity recognition, Smart Health 23 (2022), 100226, https://doi.org/10.1016/j.smhl.2021.100226.
[142] X. Li, J. Luo, R. Younes, ActivityGAN, in: Adjun. Proc. 2020 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2020 ACM Int. Symp. Wearable Comput., ACM, New York, NY, USA, 2020, pp. 249–254, https://doi.org/10.1145/3410530.3414367.
[143] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM, Visual Explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2016) 336–359, https://doi.org/10.1007/s11263-019-01228-7.
[144] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, J. Comput. Aided Mol. Des. 34 (2020) 1013–1026, https://doi.org/10.1007/s10822-020-00314-0.
[145] P.P. Ray, A review on TinyML: state-of-the-art and prospects, J. King Saud Univ. - Comput. Inf. Sci. 34 (2022) 1595–1623, https://doi.org/10.1016/j.jksuci.2021.11.019.
[146] D.L. Dutta, S. Bharali, TinyML meets IoT: a comprehensive survey, Internet of things 16 (2021) 100461, https://doi.org/10.1016/j.iot.2021.100461.
[147] A. Mukherjee, A. Bose, D.P. Chaudhuri, A. Kumar, A. Chatterjee, S.K. Ray, A. Ghosh, Edge-based human activity recognition system for smart healthcare, J. Inst. Eng. Ser. B. 103 (2022) 809–815, https://doi.org/10.1007/s40031-021-00663-w.
[148] T. Manoj, G.S. Thyagaraju, Ambient assisted living: a research on human activity recognition and vital health sign monitoring using deep learning approaches, Int. J. Innovative Technol. Explor. Eng. 8 (2019) 531–540, https://doi.org/10.35940/ijitee.F1111.0486S419.
[149] J. Arunnehru, G. Chamundeeswari, S.P. Bharathi, Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos, Procedia Comput. Sci. 133 (2018) 471–477, https://doi.org/10.1016/j.procs.2018.07.059.
[150] M. Babiker, O.O. Khalifa, K.K. Htike, A. Hassan, M. Zaharadeen, Automated daily human activity recognition for video surveillance using neural network, in: 2017 IEEE 4th Int. Conf. Smart Instrumentation, Meas. Appl., IEEE, 2017, pp. 1–5, https://doi.org/10.1109/ICSIMA.2017.8312024.
[151] F. Ma, Action recognition of dance video learning based on embedded system and computer vision image, Microprocess. Microsyst. 81 (2021), https://doi.org/10.1016/j.micpro.2020.103779, 103779.
[152] I. Lillo, J.C. Niebles, A. Soto, Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos, Image Vis Comput. 59 (2017) 63–75, https://doi.org/10.1016/j.imavis.2016.11.004.
[153] N. Islam, Y. Faheem, I.U. Din, M. Talha, M. Guizani, M. Khalil, A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services, Future Generat. Comput. Syst. 100 (2019) 569–578, https://doi.org/10.1016/j.future.2019.05.059.
[154] S. Nooruddin, M.M. Islam, F.A. Sharna, H. Alhetari, M.N. Kabir, Sensor-based fall detection systems: a review, J. Ambient Intell. Hum. Comput. 13 (2022) 2735–2751, https://doi.org/10.1007/s12652-021-03248-z.
[155] P. Vallabh, R. Malekian, Fall detection monitoring systems: a comprehensive review, J. Ambient Intell. Hum. Comput. 9 (2018) 1809–1833, https://doi.org/10.1007/s12652-017-0592-3.
[156] L. Ren, Y. Peng, Research of fall detection and fall prevention technologies: a systematic review, IEEE Access 7 (2019) 77702–77722, https://doi.org/10.1109/ACCESS.2019.2922708.
[157] P. Bet, P.C. Castro, M.A. Ponti, Fall detection and fall risk assessment in older person using wearable sensors: a systematic review, Int. J. Med. Inf. 130 (2019), 103946, https://doi.org/10.1016/j.ijmedinf.2019.08.006.
[158] A. Sathyanarayana, J. Srivastava, L. Fernandez-Luque, The science of sweet dreams: predicting sleep efficiency from wearable device data, Computer (Long. Beach. Calif). 50 (2017) 30–38, https://doi.org/10.1109/MC.2017.91.
[159] Y. Fu, J. Guo, Blood cholesterol monitoring with smartphone as miniaturized electrochemical analyzer for cardiovascular disease prevention, IEEE Trans. Biomed. Circ. Syst. 12 (2018) 784–790, https://doi.org/10.1109/TBCAS.2018.2845856.
[160] M. Panwar, D. Biswas, H. Bajaj, M. Jobges, R. Turk, K. Maharatna, A. Acharyya, Rehab-Net: Deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation, IEEE Trans. Biomed. Eng. 66 (2019) 3026–3037, https://doi.org/10.1109/TBME.2019.2899927.
[161] D. Fozoonmayeh, H.V. Le, E. Wittfoth, C. Geng, N. Ha, J. Wang, M. Vasilenko, Y. Ahn, D.M. Woodbridge, A scalable smartwatch-based medication intake detection system using distributed machine learning, J. Med. Syst. 44 (2020) 1–14, https://doi.org/10.1007/s10916-019-1518-8.
[162] S. Yamanaka, V. Moshnyaga, New Method for Medical Intake Detection by Kinect, in: 2018 IEEE 61st Int. Midwest Symp. Circuits Syst, IEEE, 2018, pp. 218–221.
[163] K. Kyritsis, C. Diou, A. Delopoulos, A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches, IEEE J. Biomed. Health Inf. 25 (2021) 22–34, https://doi.org/10.1109/JBHI.2020.2984907.
[164] K. Kyritsis, C. Diou, A. Delopoulos, End-to-end learning for measuring in-meal eating behavior from a smartwatch, in: 2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc, IEEE, 2018, pp. 5511–5514, https://doi.org/10.1109/EMBC.2018.8513627.
[165] P. Patel, B. Bhatt, B. Patel, Human body posture recognition — a survey, in: 2017 Int. Conf. Innov. Mech. Ind. Appl, IEEE, 2017, pp. 473–477, https://doi.org/10.1109/ICIMIA.2017.7975660.
[166] M.J. Cheok, Z. Omar, M.H. Jaward, A review of hand gesture and sign language recognition techniques, Int. J. Mach. Learn. Cybern. 10 (2019) 131–153, https://doi.org/10.1007/s13042-017-0705-5.
[167] A.S. Kundu, O. Mazumder, P.K. Lenka, S. Bhaumik, Hand gesture recognition based omnidirectional wheelchair control using IMU and EMG sensors, J. Intell. Rob. Syst. 91 (2018) 529–541, https://doi.org/10.1007/s10846-017-0725-0.
[168] R. Chereshnev, A. Kertész-Farkas, HuGaDB: human gait database for activity recognition from wearable inertial sensor networks, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2018, pp. 131–141, https://doi.org/10.1007/978-3-319-73013-4_12.
[169] J. Figueiredo, C.P. Santos, J.C. Moreno, Automatic recognition of gait patterns in human motor disorders using machine learning: a review, Med. Eng. Phys. 53 (2018) 1–12, https://doi.org/10.1016/j.medengphy.2017.12.006.
[170] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, Gait recognition for person re-identification, J. Supercomput. 77 (2021) 3653–3672, https://doi.org/10.1007/s11227-020-03409-5.
[171] M. Shahverdy, M. Fathy, R. Berangi, M. Sabokrou, Driver behavior detection and classification using deep convolutional neural networks, Expert Syst. Appl. 149 (2020), 113240, https://doi.org/10.1016/j.eswa.2020.113240.
[172] M. Taherisadr, P. Asnani, S. Galster, O. Dehzangi, ECG-based driver inattention identification during naturalistic driving using Mel-frequency cepstrum 2-D transform and convolutional neural networks, Smart Health 9–10 (2018) 50–61, https://doi.org/10.1016/j.smhl.2018.07.022.
[173] A.I.N. Alshbatat, S. Alhameli, S. Almazrouei, S. Alhameli, W. Almarar, Automated vision-based surveillance system to detect drowning incidents in swimming pools, Ina. Sci. Eng. Technol. Int. Conf. (ASET) (2020) 1–5, https://doi.org/10.1109/ASET48392.2020.9118248.
[174] Y. Park, Y. Jeong, C. Sohn, Suspicious behavior recognition using deep learning, J. Adv. Mil. Stud. 4 (2021) 43–59, https://doi.org/10.37944/jams.v4i1.78.
[175] M. Yang, S. Rajasegarar, S.M. Erfani, C. Leckie, Deep learning and one-class SVM based anomalous crowd detection, in: 2019 Int. Jt. Conf. Neural Networks, IEEE, 2019, pp. 1–8.
[176] G. Brunner, D. Melnyk, B. Sigfússon, R. Wattenhofer, Swimming style recognition and lap counting using a smartwatch and deep learning, Proc. - Int. Symp. Wearable Comput. ISWC (2019) 23–31, https://doi.org/10.1145/3341163.3347719.
[177] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, F.Y. Wang, Driver activity recognition for intelligent vehicles: a deep learning approach, IEEE Trans. Veh. Technol. 68 (2019) 5379–5390, https://doi.org/10.1109/TVT.2019.2908425.
[178] E. Carlson, P. Saari, B. Burger, P. Toiviainen, Dance to your own drum: identification of musical genre and individual dancer from motion capture using machine learning, J. N. Music Res. 49 (2020) 162–177, https://doi.org/10.1080/09298215.2020.1711778.
[179] F. Zhu, R. Zhu, Dance action recognition and pose estimation based on deep convolutional neural network, Trait. Du. Signal 38 (2021) 529–538, https://doi.org/10.18280/ts.380233.
[180] D. Luvizon, D. Picard, H. Tabia, Multi-task deep learning for real-time 3D human pose estimation and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2020) 2752–2764, https://doi.org/10.1109/TPAMI.2020.2976014.
[181] K. Kim, A. Jalal, M. Mahmood, Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents,

J. Electr. Eng. Technol. 14 (2019) 2567–2573, https://doi.org/10.1007/s42835-019-00278-8.
[182] M. Mlakar, M. Luštrek, Analyzing tennis game through sensor data with machine learning and multi-objective optimization, in: Proc. 2017 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2017 ACM Int. Symp, . Wearable Comput., 2017, pp. 153–156.
[183] V. Reno, N. Mosca, R. Marani, M. Nitti, T. D'Orazio, E. Stella, Convolutional neural networks based ball detection in tennis games, in: 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work., IEEE, 2018, pp. 1839–1845, https://doi.org/10.1109/CVPRW.2018.00228.

Md. Milon Islam is currently pursuing his PhD at the Centre for Pattern Analysis and Machine Intelligence in the Department of Electrical and Computer Engineering at the University of Waterloo, Canada. He received his B.Sc. and M.Sc. degree in Computer Science and Engineering (CSE) from the Khulna University of Engineering & Technology, Khulna, Bangladesh, in 2016 and 2019 respectively, where he is currently working as an Assistant Professor (on leave) in CSE department. In 2017, he joined the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, as a Lecturer. He has published several research papers in peer reviewed journals, book chapters and conferences. In addition, he is reviewer of several reputed journals and conferences. His research interests include machine learning and its application, deep learning, intelligent systems design, health informatics, Internet of Things (IoT), and to solve real life problems with the concept of computer science.

Sheikh Nooruddin is currently pursuing his Master of Applied Science (MASc) degree in Electrical and Computer Engineering (ECE) with a specialization in Pattern Analysis and Machine Intelligence (PAMI) at the University of Waterloo, Ontario, Canada. He received his B.Sc. Eng. in Computer Science and Engineering degree from Khulna University of Engineering & Technology, Bangladesh in 2020. His research interests include computer vision, artificial intelligence, human activity recognition, Internet of things, and medical signal processing. He has authored and co-authored multiple journals and conference articles published by reputed peer-reviewed international publishers including IEEE, Elsevier, and Springer.

Fakhri Karray is a Professor and Provost of the Mohamed Bin Zayed University of AI in the UAE. He has served as the University Research Chair Professor in Electrical and Computer Engineering and the founding co-director of the Institute of Artificial Intelligence at the University of Waterloo. He holds the Loblaw's Research Chair in Artificial Intelligence. Dr. Karray's research work spans the areas of intelligent systems and operational artificial intelligence as applied to smart devices and man machine interaction systems through speech, gesture, and natural language. He has authored extensively in these areas and has disseminated his work in journals, conference proceedings, and textbooks. He is the co-author of two dozen US patents, has chaired/co-chaired several international conferences in his area of expertise and has served as keynote/plenary speaker on numerous occasions. He has served as the associate editor/guest editor for a variety of leading journals in the field, including the IEEE Transactions on Cybernetics, the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Mechatronics, the IEEE Computational Intelligence Magazine. He has served as the University of Waterloo's Academic Advisor for Amazon's Alexa Fund Fellowship Program and is a Fellow of the Canadian Academy of Engineering, a Fellow of the Engineering Institute of Canada and a Fellow of the IEEE.

Ghulam Muhammad (Senior Member, IEEE) is a professor in the Department of Computer Engineering, College of Computer and Information Sciences at King Saud University (KSU), Riyadh, Saudi Arabia. Prof. Ghulam received his Ph.D. degree in Electronic and Information Engineering from Toyohashi University and Technology, Japan in 2006, M.S. degree from the same university in 2003. He received his B.S. degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology in 1997. He was a recipient of the Japan Society for Promotion and Science (JSPS) fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan. His research interests include signal processing, machine learning, IoTs, medical signal and image analysis, AI, and biometrics. Prof. Ghulam has authored and co-authored more than 300 publications including IEEE / ACM / Springer / Elsevier journals, and flagship conference papers. He owns two U.S. patents. He received the best faculty award of Computer Engineering department at KSU during 2014-2015. He has supervised more than 15 Ph.D. and Master Theses.