An Overview of Deep Learning Methods in the Internet of Things Technology in regular life

[1]Seyed Ebrahim Dashti, [2]Ahmed Rahman Abdulzahra Al-Obaidi, [3]Saba Atiyah Mashaan

[1]Department of Computer Engineering, Jahrom Branch, Islamic Azad University

[2,3]Department of Computer Engineering, Shiraz Branch, Islamic Azad University

Abstract:

Large volumes of data are generated daily as a result of the extensive usage of Internet of Things (IoT) technology in indoor daily life. dependable methods for data analysis are necessary to make efficient use of this data. Recent advances in deep learning (DL) make it easier to handle and learn from enormous amounts of IoT data, allowing for a quick and competent understanding of the fundamentals of many IoT applications in intelligent indoor settings. The current literature on the usage of DL for various indoor IoT applications is summarized in this paper. Our objective is to provide knowledge on how to apply deep learning techniques from many angles to create better two separate indoor IoT application areas: indoor localization/tracking and activity detection. One important objective is to seamlessly combine the two fields of IoT and deep learning, which will lead to a variety of creative approaches for indoor IoT applications including robots, smart home automation, health monitoring, etc. Additionally, from a comparison of technical research in the three aforementioned categories, we develop a thematic classification. To increase the effectiveness of indoor IoT applications and to encourage and inspire further advancement in this exciting field of study, we conclude by proposing and discussing various problems, difficulties, and new directions to apply deep learning.

Keywords: Deep learning, Internet of Things, IoT,Overview,

DOI: 10.24297/j.cims.2023.5.1

1. Introduction

Conditional logic to neural networks is only a few example of the many types of automated decision-making that artificial intelligence (AI) encompasses. Machine learning (ML), a branch of artificial intelligence, is used to make predictions or decisions. Deep learning (DL) is a term used to describe a subset of machine learning methods that employ deep neural networks (DNNs). Research publications on artificial intelligence (AI) now account for 3% of journal articles and 9% of conference papers published in the last two decades. [1]. Most AI research typically focuses on developing algorithms and optimization techniques, with a focus on checking the accuracy of high-performance models. artificial intelligence, machine learning,and deep learning are widely used in data-rich sectors in addition to academic study. With varied degrees of success, the same sectors have built goods and services on top of the AI backend. The necessity to evaluate and enhance algorithmic models for conversion into end-user requirements grows along with the industry's adoption of sophisticated models.The goal of creating smart building systems is to employ technology to enhance indoor living conditions for people [1]. To enhance residents' interior quality of life, we have developed a number of smart home applications, including B. interior device remote control, indoor fire detection, gas leak detection, power saving, elderly monitoring, childcare, and gesture control [2]. Together, Multiple purposes offer the use of smart indoor systems to simplify everyday tasks, lessen human effort, and highlight anxieties about problematic or uncomfortable situations at home. Recently, We now have access to a vast amount of information on individuals, particularly indoor data, because of the wide adoption and significant advancements in sensing techniques, Internet of Things (IoT) technologies, and communication technologies. To provide a variety of indoor services or applications that improve people's lives, this enormous diversity of data can be gathered, cleaned up, and evaluated.[3]. The creation of intelligent Internet of Things (IoT) applications has enhanced the viability of creating intelligent systems that enhance people's interior quality of life. Finally, it is generally agreed that the concept of intelligent indoor data analysis may be summed up as a five-step workflow: definition of the problem, data collection, preparation, analysis, and service provision. This is a little variation from earlier studies.

1.1 Innovation

This research helps with that by offering a thorough review of the function of DL techniques in indoor IoT use cases based on the following criteria:

- For user-centric Applications for the Internet of Things (IoT) in smart indoor settings, this paper presents a thorough analysis and classification of recent breakthroughs in deep learning techniques. First, we group current research into categories based on how dependent it is on particular devices (sensor technology). Second, the underlying learning technique automatically categorizes and compares DL models. Finally, we offer a taxonomy to organize current literature according to application domains.
- This research also provides a tabular summary of publicly accessible benchmarks and datasets, such as visual, sensor-based, radio frequency, and other data. The primary objective aims to educate scholars about the freely available data that can be utilized to

test and evaluate newly developed deep learning methods for indoor IoT applications with a human-centered focus.

- Using a one-to-one comparison table, we explore and assess various deep-learning strategies for human-centric indoor applications. The proposed deep learning models, accuracy, application, system design, and data used for these methods are contrasted.
- This study discusses current shortcomings, challenges, and future opportunities for exploring deep learning techniques to improve improving the efficacy and productivity of IoT applications with a centered people focus aimed at improving people's quality of life indoors.

1.2 HCML's (Human-Centered Machine Learning) rise

For more than ten years, HCML nomenclature has been utilized in publications. Balasubramanian et al. assert that their study on human-in-the-loop ML systems is what gave rise to the term "human-centric machine learning algorithms" in an aid. However, the use of the term HCML surged and gained popularity in the middle of the 2010s, when the modern deep-learning period began. The possibilities, constraints, and internal workings of AI are still not fully understood by the typical user. As a result, [3,4] consumers express worries about AI systems' ability to explain things, user experience (UX), user privacy, security, and dependability. The area of Human-Centered Machine Learning (HCML) has developed in response to the necessity to solve these issues. The HCML area aims to enhance user-centric ML system development by recognizing that algorithmic optimization and cutting-edge neural network topologies alone cannot address usability and acceptance problems. Today, HCML words are mentioned in both formal and informal contexts, including AI publications[5], workshops, conferences, blogs, and articles from businesses with an AI focus. HCML emerged as a topic of research to investigate approaches for integrating machine learning systems with human objectives, circumstances, worries, and working styles. In addition to the many institutions that research the user experience elements of AI, many terminologies and acronyms have also been developed. Institutions now refer to ML and AI as "human-centered machine learning" (HCML) or "human-centered artificial intelligence. However, the rationale for all of these terms remains the same, which is to develop usable and adaptable "human-in-the-loop" machine learning systems.

2. Related surveys

This part analyzes current studies of indoor IoT applications, covering both applications that depend on specific devices and those that don't.

2.1 Surveys on device-dependent Approaches

Various device-dependent approaches have been developed for different types of IoT applications, especially those based on camera data or sensor data. Li et al. [14] provide a comprehensive overview of multiple computational techniques for physical activity detection and related applications in a smart IoT environment. They mainly discuss the analysis and fusion of sensory data (sights or sensors) and offer some insights into the challenges and possibilities of collective activity recognition. Abu Hamed et al. [15] reviewed about 140 studies on persistent human authentication techniques, classifying them into six interactive and physical biometric categories, including gesture, gait, speech, movement, keystroke dynamics, and multimodality. They also compare related research based on sensor, modality, algorithm, and user data. The authors also discuss intuitions and challenges of current biometrics that can be addressed in future work. Dang et al. [12] examined and analyzed research on human AR methods and showed their respective advantages and disadvantages, classifying AR methods into two categories, namely H. methodologies using sensors and vision that rely on data aggregation, feature extraction, preprocessing techniques, and training techniques. The writers also discuss group activities, gestures, actions, and human activities at various levels of human-object (HO) and human-human (H-H) interaction. In their discussion of fusion-based localization techniques, Guo et al. [16] covered heterogeneous, homogeneous, and hybrid systems as well as sources from diverse network frameworks. Device-dependent techniques, however, are obtrusive and difficult to employ since they necessitate a connection or co-location between the user and the device. They are severely affected by environmental barriers.

2.2 Device-Independent Approaches Surveys

Device-independent sensor technologies have been created to make it possible to design indoor IoT applications without relying on connected devices or security cameras, thus overcoming the drawbacks of the aforementioned device-dependent techniques. Hussein et al.'s [13] evaluation of device-free techniques for identifying various indoor human activity categories, such as motion-based, and activities based on interaction. They provide a taxonomy as well with ten distinct subcategories for this task of activity identification. Using various kinds of fingerprints, Zhu et al. [17] investigated ML and smart indoor localization techniques and unveiled a novel smart localization framework. The primary problems in designing intelligent localization for the actual world are also covered, along with suggestions for future advancements and solutions. In IoT contexts, Alam [8] offers a thorough discussion of non-RF-based techniques for device-independent indoor localization. The author covers studies that use light, infrared, physical excitation, and electric-field detection, and then they talk about each method's major drawbacks and future research opportunities. Research on unusual behaviors in geriatric care in indoor IoT contexts is outlined by Deep et al. [18], who emphasize the stability, non-intrusiveness, and sociability of dense perception-based responses to contextual fluctuations. The authors also discuss important problems and links between conduct that is abnormal and human activities.Nimar et al. [19] examine several various algorithms and offer a more thorough examination of DL for human-centric RF-based sensing in their thorough assessment and categorization of DL research for RF-based human sensing. They also examined 20 published benchmarks for radio emissions that can be used to detect human activities. They are both together. The capture, identification, and detection of channel status information (CSI) from commercial WiFi equipment is one example of a recent breakthrough in WiFi vision tasks that was studied by [2].

They emphasize the applicability of these tasks in nine essential IoT contexts, such as WiFi image processing, vital sign monitoring, indoor localization, gesture and gait analysis, standard AR, fall detection, and person recognition are some examples of related technologies. Wireless sensor techniques were investigated by Liu et al. [20] in the context of their fundamental forerunners, methodology, and system designs. The utilization of wireless signals to streamline the design of various IoT applications, such as interior location anomaly detection, room occupancy monitoring, commonplace AR, gesture recognition, vital sign monitoring, and person recognition, is then covered. They also discuss the potential for future human-centric applications utilizing wireless signals. According to Thariq Ahmed [21], there are two types of gesture recognition techniques: model-based and learning-based. These techniques are used in contexts with device-independent sensing. Additionally, they go into data preparation, using feature engineering and classification models, and performance-affecting environmental factors. To assess overall performance and enhance useful human recognition methods, Zhang [5] also looked into ML- and DL-based wireless sensing for people detection using RGB/depth pictures and radar data.

3. Deep Learning Techniques

To enhance IoT operations and services, big data research has focused on IoT as one of its major producers. IoT big data research also demonstrates that it can benefit society. IoT data differs from big data in its entirety. To study the requirements of IoT data analysis, it can look at the characteristics of IoT data and how it differs from conventional big data[25].

The benefits of deep learning over conventional machine learning techniques can be discussed here, with an emphasis on the benefits of deep learning in IoT applications. DL has a better capacity to generalize the dynamic relationships of vast volumes of raw data in diverse IoT applications than traditional ML methods[27,28]. Deep learning models are likely to perform better on big data, whereas conventional learning models can easily become overwhelmed when dealing with enormous amounts of data. The ability to process data generally depends on the depth of the learning model and different architectures, including convolutional architectures. Deep learning is an end-to-end method that can learn how to create useful features from unstructured data without relying on labor- and time-intensive manual processes. In comparison to other conventional ML techniques, DL models have improved in sophistication recently.

An advanced multi-layer neural network learning algorithm is called deep learning. It advanced artificial intelligence and human-computer interaction while revolutionizing the idea of machine learning. On the MNIST database and the real-world handwritten character database, they gave the CNN and DBN scores, with 99.28% and 98.12% accuracy, respectively[29]. Researchers here assert that MIA may be employed in semi-white-box scenarios where system model structure and parameters are known but no user data information is available, and even consider it a severe concern due to its complicated structure and a vast variety of registered user data. Threats verified using facial recognition technology based on deep learning. This essay investigates how a power plant affects GEP throughout its lifetime[31]. Time series forecasting also makes extensive use of deep learning-based approaches.

Large volumes of unstructured information can be handled by the DL's potent knowledge expansion methods[32]. Large-scale data management and computationally demanding tasks like speech recognition, visual pattern recognition, and analytics are best handled by these technologies. The model training cycle in DL is known to be time-consuming and demands high computational abilities, which has been one of the main obstacles in the past. Deep learning tasks that demand more CPU power are frequently executed on effective GPUs. As a result, DL has gained popularity as a method of data processing and modeling in the age of big data. There aren't many layers with distinct attributes in the DL approach[27]. In DL, features are automatically estimated, hence no feature extraction or computation is necessary before the use of such a procedure. Furthermore, numerous network architectures have been added as a result of DL advancements. The authors' project's [33,34] objectives are to analyze biomarkers to distinguish ischemic stroke patients from healthy people and to quantify EEG signals to better understand task-related neurological impairment induced by stroke.

In comparison to conventional ML techniques, DL models typically provide two key advantages throughout the training and prediction phases. They initially reduce the need for human training before eliminating any elements that might not be visible to human vision. DL techniques can boost precision as well. Like conventional M, DL[21]. Models with unlabeled data go under the category of unsupervised learning, while models with labeled data fall under the category of supervised learning.

3.1 Supervised Learning

The identified training set contains the system model for supervised learning. In supervised learning, the backpropagation method is the main strategy utilized.

3.1.1 Recurrent Neural Networks (RNNs)

RNN is a discriminative classification technique that works best with time series and sequence data. To evaluate the input sequence for particular tasks, the estimate uses multiple previous tests in addition to the categorization of a single test. Since a feeder neural network does not rely on input and output layers, it is inappropriate for these applications. The input to the RNN comprises both the current sample and samples that have already been observed[23]. The output of stage m-1 has an impact on stage m's. Each neuron contains a feedback loop that acts as both an input and the subsequent output. This procedure shows that each neuron in the RNN[5] has an internal memory for storing data estimations from the preceding layer. Even though there are neural loops, we can't use raw backdrop in this situation since it relies on the derivative of the weight loss from the preceding layer, even though the RNN doesn't have a stacked layer model. We create a network of feeders through time by utilizing the "unrolled RNN" method, which is at the core of backpropagation through time (BPTT).

Due to the predominance of gradient issues and long-term dependencies, RNNs can only go back a short distance. To choose what to store in past and present memory, new techniques have been developed, such as GRU (Gated Recurrent Unit) and LSTM (Long Short Term

Memory)[19]. RNNs were created to address issues that required sequential solutions, such as time series data of various lengths and text or language. RNNs can be used for a variety of tasks, such as determining personal mobility patterns, determining household usage, and recognizing driving behavior in smart cars. RNNs are therefore mostly employed in the natural language processing (NLP) industry.

3.1.2 Long Short-Term Memory (LSTM)

A discriminative technique called LSTM is capable of processing time-stamped, sequential, and long-term dependant data. An RNN variant called LSTM picks up on order dependencies in sequence estimation. LSTMs compute a value between 0 and 1 using their unit gates, each of which is predicated on its input. Four gates are built into each neuron to store data; for instance, these gates regulate access to memory cells and shield them from unwanted input. The neuron writes its data to itself when the gate forgets to function; otherwise, it sends a 0 to indicate that it has forgotten its previous data. Other connected neurons can write to and read from the read-write gate when it is fixed at 1. Calculations performed on stored memory cells are not corrupted over time by knowing which LSTM data to fetch. BPTT is a popular technique for error reduction through network training. When the data has long-term dependencies, LSTM models outperform RNN models[38].

LSTMs are often developed versions of RNNs. Numerous LSTM techniques have been put out using the original network as a base. Sequence prediction and sequence labeling tasks have been accomplished well using LSTMs and conventional RNNs[58]. In both context-sensitive (CS) and context-free (CF) languages, these models outperform RNNs. Modern machine translation and effective speech recognition are provided by LSTMs for connected models with small sizes. On a single multi-core machine, LSTM networks are not appropriate for big networks.

3.1.3 Convolutional Neural Networks (CNNs)

CNN is a discriminative method that works well for recognizing and differentiating pictures. Input, output, and a few hidden layers make up CNN. In CNN designs, hidden layers can be subsample layers, pooling layers, convolutional layers, pooling, fully connected (FC), or nonlinear layers. The primary iteration of FC is CNN. The connections between each neuron in each layer are complete. The data was therefore overwritten by FC[22].

The translation invariance property of DNNs with deep-layer interactions makes them challenging to train and poorly tested on tasks requiring vision. CNN can resolve these issues with the assistance of the aforementioned features. Using 2D input, CNNs can extract high-quality characters from a variety of hidden layers, including audio or images. The core of CNNs is convolutional layers, which comprise filters with the same input form but lower dimensions. Complex networks can incorporate global or local pooling layers to streamline low-level data processing by reducing data dimensionality by integrating neuron cluster outputs into neurons in the following layer[78,79]. Since the activation features and final convolutions cover their inputs and outputs, RELU layers are typically activation functions supplemented by extra convolutions, such as pooling layers, FC layers, and hidden layers.

3.1.4 Transformer-Based Deep Neural Networks

Transformers are sequence-to-sequence neural network designs that use self-awareness techniques to detect global relationships in the context of deep learning. Many natural language processing (NLP) academics were interested in it because the transformer was created with sequence data as an input. Bidirectional Encoder Representation (BERT) from Transformers is one of the most effective Transformer-based models, attaining state-of-the-art performance in numerous NLP tasks. Transformers have also recently gained wider acceptance in the field of computer vision. Dosovitskiy et al. created a categorization system for transformer images utilizing image portions as input. Suggest.

Detection Transformer (DETR)[115] is a successful effort for an end-to-end object detection framework based on Transformer. By eliminating various manually created components that encode past information, such as B. Spatial anchors in place of maximum suppression, DETR streamlines the object detection workflow. As a result, deep neural networks built on Transformers are also viable methods for tackling AI-related problems including NLP and computer vision-related disciplines.

3.2 Unsupervised Learning

To deal with vast amounts of unlabeled data, unsupervised learning must be employed in addition to traditional learning techniques. To initialize, duplicate back, and alter globally during training, stacked restricted Boltzmann machines (RBMs) or stacked auto-encoders can be used.

3.2.1. Autoencoder (AE)

With the same number of input and output units, AE is a generative technique that may be modified to extract features and minimize size. One or more hidden layers are connected to these input and output layers. An autoencoder is a neural network that is set up to duplicate its input to its output. To make the input visible, the code level is private (hidden). An encryption encoder that maps the code input and a decoder that maps the code to decrypt the original input make up the two primary components of this layer[24]. The autoencoder has a function for reducing input and output errors. As a result of their function as input generators on the output layer, AEs are mostly employed for diagnosis and error identification. It will highlight a variety of IoTapplications[25]. Sparse autoencoders, denoisingautoencoders, and systolic autoencoders are examples of AE variants.

3.2.2 Restricted Boltzmann Machines (RBMs)

RBM is a generative method that can handle different types of data and can be used for data classification, dimensionality reduction, feature extraction, etc. RBMs are deep random networks that are probabilistic graphical models. The ability of an RBM's neurons to build a bipartite graph is constrained by the Boltzmann version. A pair of nodes in a hidden group and a visible group may have a symmetric relationship. There is no relationship between nodes belonging to the same group, nevertheless. Additionally, biasing devices are attached to all neurons, both visible and secret (hidden). To create the DNN, it could be essential to store the RBM. They serve as the DBN network's skeleton as well. In particular, RBMs can be stacked to create DBNs, or related deep gradient descent and backpropagation networks can be adjusted. Optimizing the product for all probabilities of visible units is the aim of RBM training. For measuring latent parameters, which are then utilized to reverse-flow reconstruct data inputs, RBM has traits with AE.

3.2.3 Deep Belief Networks (DBNs)

DBN is a generative technique that can handle many data types. A DBN can be compared to a collection of straightforward unsupervised networks (like RBM and AE), where each sub-networks hidden layer serves as the visible layer of the following sub-network. Such a network does not have connections within the layers, only between them[27]. Layer by layer, DBNs can also be trained greedily. Due to this interaction, the "bottom" layers carry out a quick, unsupervised training process during which each sub-network is subjected to opposite divergences. DBN training is carried out layer by layer, with each layer being treated as an RBM that has been trained on top of earlier trained layers. DBN can therefore be quick and effective in DL approaches.

3.3 Deep semi-supervised learning models

The scope of semi-supervised DL models includes models designed to utilize instances of unlabeled and labeled data during training. For example, Under arbitrary perturbations (such as scaling, rotation, translation, flipping, or random shaking), an efficient DL model must generate smooth and soft estimates of GANs[30]. Current semi-supervised models can be viewed from two different perspectives, generative models and teacher-student models.

3.4 Generative modelsof the semi-supervised

AEs, RBMs, DBF, and GANs could be obtained from the equivalent unsupervised DL models. then treats them as a subset of the first K authentic classes to determine the distribution of unlabeled observations. Additionally, Utilizing the latent representation encoded from the labeled and unlabeled portion of the training data, the semi-supervised AE creates a classifier for forecasting[31].

3.4.1 Teacher–student models

TheTeacher–Student models are regarded as the type of semi-supervised models that have been realizing great success in recent years. During training, The parameters of a student network are managed by the estimated labels. To increase the student network's capacity to classify the unlabeled observations, the consistency between the teacher and the students needs to be enhanced.

4. Data collection and benchmarks

Open-source datasets are becoming increasingly necessary for the DL research community to promote reproducibility and quicken the pace of research output. The final objective is to offer multiple benchmarks to quickly experiment with and compare the effectiveness of DL models from multiple, independent studies. Aggregating and annotating human-centric datasets is a challenging task, regrettably. Due to privacy restrictions, particularly for indoor data, publicly available data is not always accessible, and preserving records might be expensive. Access to community/shared datasets may be perfect for expediting deep learning research in the future, even if the majority of academics are now gathering their datasets in the lab to test out proposed deep learning algorithms. Due to the current scarcity of data for these applications, this section summarizes the publicly accessible benchmarks for several indoor IoT applications in depth. This may help future research exploration.

4.1 Vision datasets

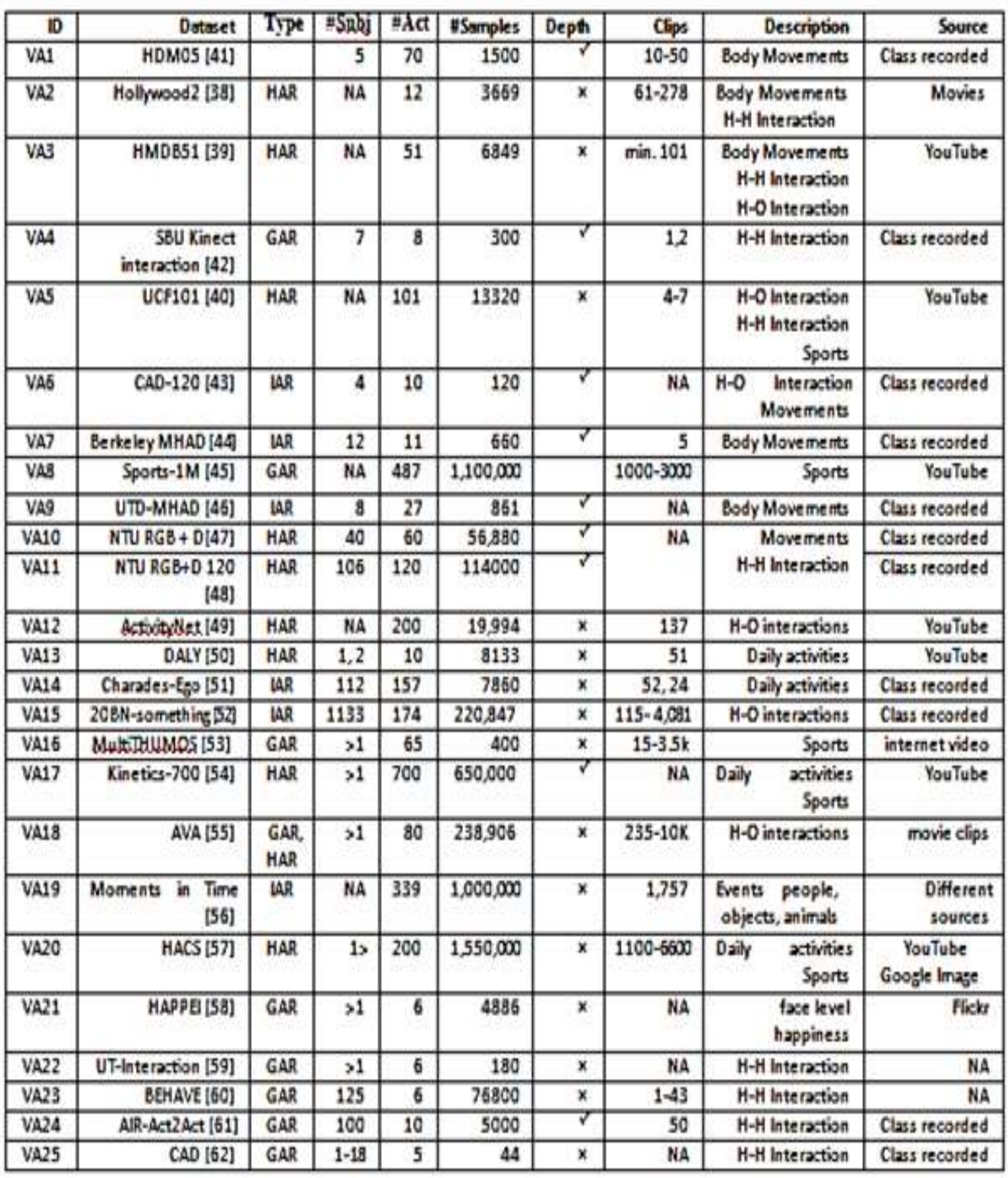
The vast improvement and wide applicability of computer vision models have drawn researchers' attention to the development of innovative AR methods based on vision datasets. Standard cameras may record RGB images, which are comprised of blue channels andred-green, in a visible continuum[38], [39], [40]. Still, cameras often have a small field of view, tend to be calibrated, and are greatly influenced by their surroundings, such as walls, lighting, and other environmental factors. With the development of depth sensors and range vision techniques, learning algorithms will be better able to distinguish human actions and actions. Still, the limited resolution of RGB-D data results in images that are noisy with silent empathy and are easily corrupted by light-sensitive and translucent materials[41], [42].

4.2 Sensory data

According to the sensor type, the present perception data can be split into four categories: environment sensors (AS), wearable sensors (WS), object sensors (OS), and mixed sensors. Additionally, it can be shown that most sensory datasets now in use only take into account the activity of a single subject, with very few taking into account the activity of several subjects or groups. Notably, the magnetometer, gyroscope, and accelerometer WISDM datasets—which are frequently used as benchmark datasets to assess sensor-dependent deep learning techniques—are incorporated into the majority of the existing benchmarks employing WS. Additionally, it can be shown that the majority of current sensory datasets are compiled by recording the behaviors of specific individuals.However, humans find it uncomfortable and potentially uncomfortable to carry a device or certain types of sensors, which makes it challenging to design ubiquitous IoT applications, especially indoor ones.

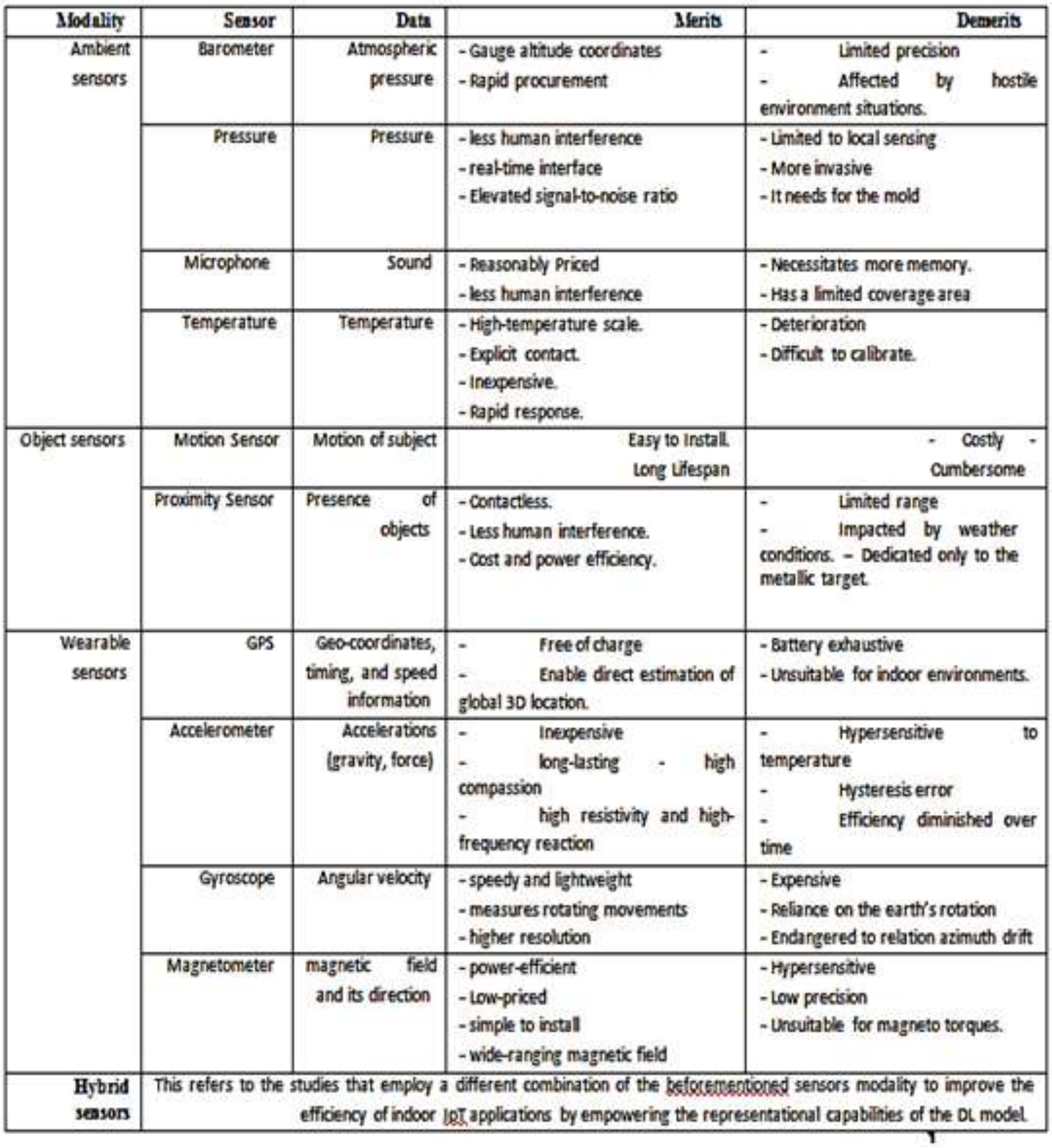
According to the sensor type, the present perception data can be split into four categories: environment sensors (AS), object sensors (OS), wearable sensors (WS), and hybrid sensors. Additionally, it can be noted that few sensory datasets now in use take several people or groups into account when aggregating sensory data[65]. It's important to note that the majority of contemporary benchmarks, including magnetometers, gyroscopes, and accelerometers, are combined utilizing WS. When assessing sensor-dependent deep learning techniques, the WISDM dataset is frequently utilized as the reference dataset. Additionally, it can be shown that the majority of current sensory datasets are compiled by recording the behaviors of specific individuals. Sadly, humans find it uncomfortable and potentially uncomfortable to carry a device or certain types of sensors, which makes it challenging to design ubiquitous IoT applications, especially indoor ones.

Table I. Summary of datasets for recognizing indoor activities using vision.



ID	Dataset	Type	#Subj	#Act	#Samples	Depth	Clips	Description	Source
VA1	HDM05 [41]		5	70	1500	✓	10-50	Body Movements	Class recorded
VA2	Hollywood2 [38]	HAR	NA	12	3669	×	61-278	Body Movements H-H Interaction	Movies
VA3	HMDB51 [39]	HAR	NA	51	6849	×	min. 101	Body Movements H-H Interaction H-O Interaction	YouTube
VA4	SBU Kinect interaction [42]	GAR	7	8	300	✓	1,2	H-H Interaction	Class recorded
VA5	UCF101 [40]	HAR	NA	101	13320	×	4-7	H-O Interaction H-H Interaction Sports	YouTube
VA6	CAD-120 [43]	IAR	4	10	120	✓	NA	H-O Interaction Movements	Class recorded
VA7	Berkeley MHAD [44]	IAR	12	11	660	✓	5	Body Movements	Class recorded
VA8	Sports-1M [45]	GAR	NA	487	1,100,000		1000-3000	Sports	YouTube
VA9	UTD-MHAD [46]	IAR	8	27	861	✓	NA	Body Movements	Class recorded
VA10	NTU RGB + D[47]	HAR	40	60	56,880	✓	NA	Movements H-H Interaction	Class recorded
VA11	NTU RGB+D 120 [48]	HAR	106	120	114000	✓			Class recorded
VA12	ActivityNet [49]	HAR	NA	200	19,994	×	137	H-O interactions	YouTube
VA13	DALY [50]	HAR	1, 2	10	8133	×	51	Daily activities	YouTube
VA14	Charades-Ego [51]	IAR	112	157	7860	×	52, 24	Daily activities	Class recorded
VA15	20BN-something [52]	IAR	1133	174	220,847	×	115-4,081	H-O interactions	Class recorded
VA16	MultiTHUMOS [53]	GAR	>1	65	400	×	15-3.5k	Sports	internet video
VA17	Kinetics-700 [54]	HAR	>1	700	650,000	✓	NA	Daily activities Sports	YouTube
VA18	AVA [55]	GAR, HAR	>1	80	238,906	×	235-10K	H-O interactions	movie clips
VA19	Moments in Time [56]	IAR	NA	339	1,000,000	×	1,757	Events people, objects, animals	Different sources
VA20	HACS [57]	HAR	1>	200	1,550,000	×	1100-6600	Daily activities Sports	YouTube Google Image
VA21	HAPPEI [58]	GAR	>1	6	4886	×	NA	face level happiness	Flickr
VA22	UT-Interaction [59]	GAR	>1	6	180	×	NA	H-H Interaction	NA
VA23	BEHAVE [60]	GAR	125	6	76800	×	1-43	H-H Interaction	NA
VA24	AIR-Act2Act [61]	GAR	100	10	5000	✓	50	H-H Interaction	Class recorded
VA25	CAD [62]	GAR	1-18	5	44	×	NA	H-H Interaction	Class recorded

Table II.A summary of the traits, benefits, and drawbacks of the various indoor sensor categories.

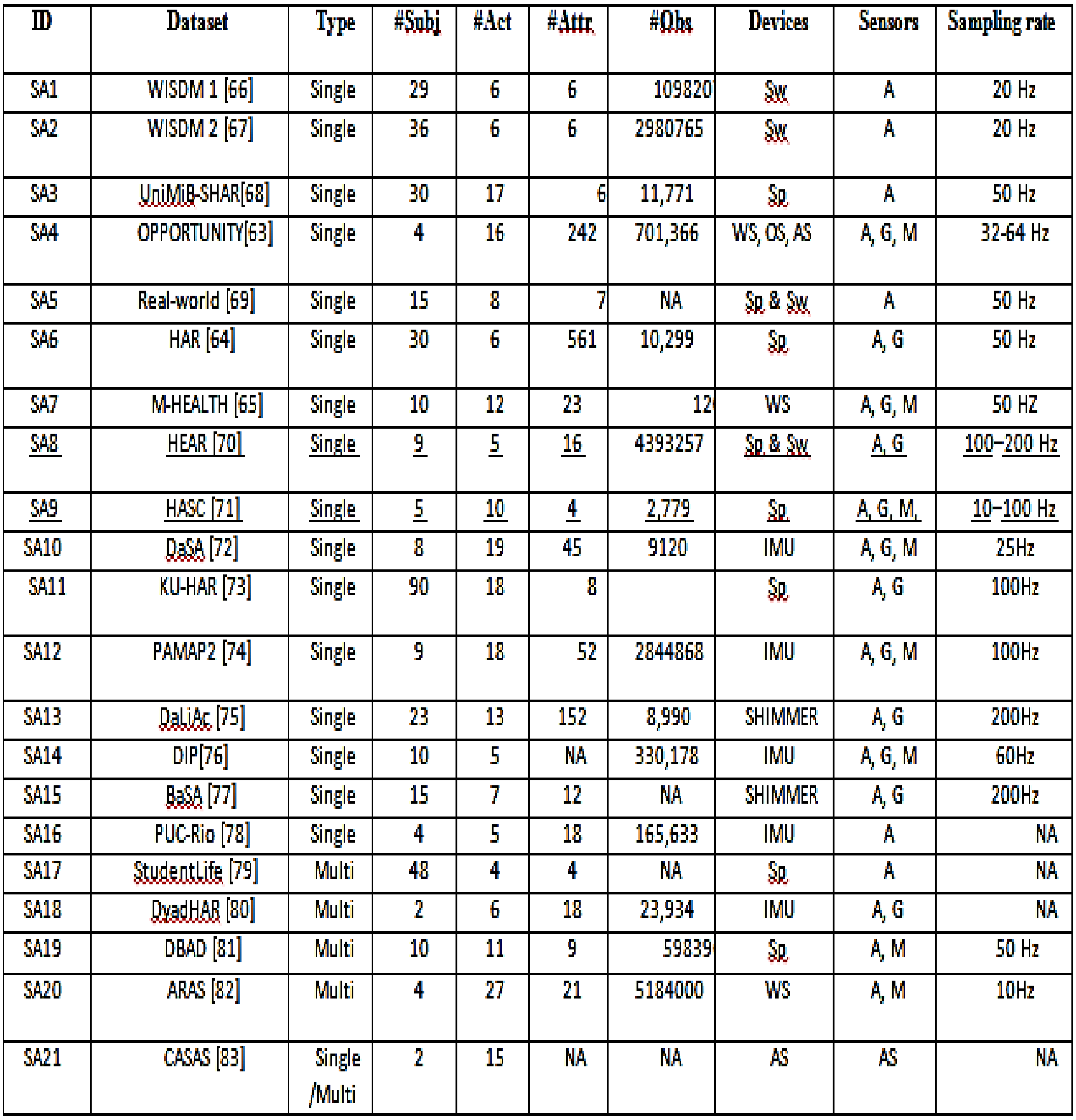


Modality	Sensor	Data	Merits	Demerits
Ambient sensors	Barometer	Atmospheric pressure	- Gauge altitude coordinates - Rapid procurement	- Limited precision - Affected by hostile environment situations.
	Pressure	Pressure	- less human interference - real-time interface - Elevated signal-to-noise ratio	- Limited to local sensing - More invasive - It needs for the mold
	Microphone	Sound	- Reasonably Priced - less human interference	- Necessitates more memory. - Has a limited coverage area
	Temperature	Temperature	- High-temperature scale. - Explicit contact. - Inexpensive. - Rapid response.	- Deterioration - Difficult to calibrate.
Object sensors	Motion Sensor	Motion of subject	Easy to Install. Long Lifespan	- Costly - Cumbersome
	Proximity Sensor	Presence of objects	- Contactless. - Less human interference. - Cost and power efficiency.	- Limited range - Impacted by weather conditions. – Dedicated only to the metallic target.
Wearable sensors	GPS	Geo-coordinates, timing, and speed information	- Free of charge - Enable direct estimation of global 3D location.	- Battery exhaustive - Unsuitable for indoor environments.
	Accelerometer	Accelerations (gravity, force)	- Inexpensive - long-lasting - high compassion - high resistivity and high-frequency reaction	- Hypersensitive to temperature - Hysteresis error - Efficiency diminished over time
	Gyroscope	Angular velocity	- speedy and lightweight - measures rotating movements - higher resolution	- Expensive - Reliance on the earth's rotation - Endangered to relation azimuth drift
	Magnetometer	magnetic field and its direction	- power-efficient - Low-priced - simple to install - wide-ranging magnetic field	- Hypersensitive - Low precision - Unsuitable for magneto torques.
Hybrid sensors	This refers to the studies that employ a different combination of the beforementioned sensors modality to improve the efficiency of indoor IoT applications by empowering the representational capabilities of the DL model.			

4.3 Radio frequency data

Due to their contactless, non-LOS, and privacy-preserving qualities, RF waves have been used by researchers to develop smart IoT applications. According to the chosen communication technology, three basic categories can be applied to RF data. With its various benefits, A low-cost communication technique called Radio Frequency Identification (RFID) does away with the need for device extras like sensors. Instead, the latter tags store energy from a nearby RFID reader that reads RF signals and analyzes them to give radar-based information collection. Radar is a type of technique for active sensing using RF waves that are transmitted and subsequently altered by the target. Radar technology for indoor applications is becoming more popular in modulated form. Continuous wave radar (CW) and ultra-wideband radar (UWB) are the two types of radars now available for this use.

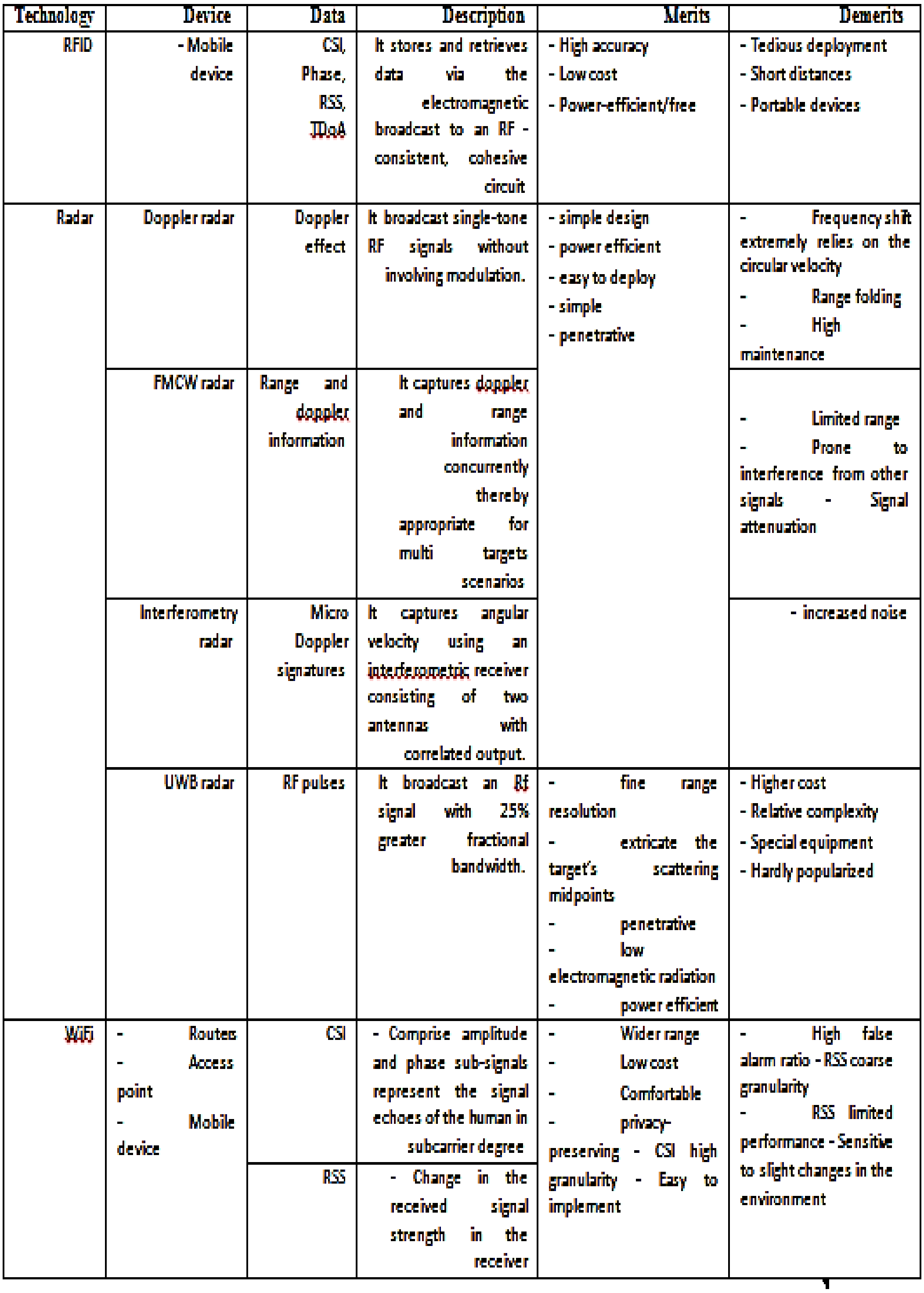
Table III. Summary of indoor activity identification datasets using sensors



ID	Dataset	Type	#Subj	#Act	#Attr	#Obs	Devices	Sensors	Sampling rate
SA1	WISDM 1 [66]	Single	29	6	6	109820	Sw	A	20 Hz
SA2	WISDM 2 [67]	Single	36	6	6	2980765	Sw	A	20 Hz
SA3	UniMiB-SHAR[68]	Single	30	17	6	11,771	Sp	A	50 Hz
SA4	OPPORTUNITY[63]	Single	4	16	242	701,366	WS, OS, AS	A, G, M	32-64 Hz
SA5	Real-world [69]	Single	15	8	7	NA	Sp & Sw	A	50 Hz
SA6	HAR [64]	Single	30	6	561	10,299	Sp	A, G	50 Hz
SA7	M-HEALTH [65]	Single	10	12	23	12	WS	A, G, M	50 HZ
SA8	HEAR [70]	Single	9	5	16	4393257	Sp & Sw	A, G	100–200 Hz
SA9	HASC [71]	Single	5	10	4	2,779	Sp	A, G, M,	10–100 Hz
SA10	DaSA [72]	Single	8	19	45	9120	IMU	A, G, M	25Hz
SA11	KU-HAR [73]	Single	90	18	8		Sp	A, G	100Hz
SA12	PAMAP2 [74]	Single	9	18	52	2844868	IMU	A, G, M	100Hz
SA13	DaLiAc [75]	Single	23	13	152	8,990	SHIMMER	A, G	200Hz
SA14	DIP[76]	Single	10	5	NA	330,178	IMU	A, G, M	60Hz
SA15	BaSA [77]	Single	15	7	12	NA	SHIMMER	A, G	200Hz
SA16	PUC-Rio [78]	Single	4	5	18	165,633	IMU	A	NA
SA17	StudentLife [79]	Multi	48	4	4	NA	Sp	A	NA
SA18	DyadHAR [80]	Multi	2	6	18	23,934	IMU	A, G	NA
SA19	DBAD [81]	Multi	10	11	9	59839	Sp	A, M	50 Hz
SA20	ARAS [82]	Multi	4	27	21	5184000	WS	A, M	10Hz
SA21	CASAS [83]	Single /Multi	2	15	NA	NA	AS	AS	NA

The signal from a target (a human) in the signal route is tempered by the notoriously consistent frequency CW ratio signal that CW radars broadcast. Doppler radar, interferometric radar, and radar. The difficulty in assessing and contrasting various DL studies is revealed by the HF data's sensitivity to experimental circumstances and equipment configuration. With a subject population spanning from 1 to 95, the current dataset was dominated by daily activity, gesture, and gait analyses. Between 1 and 7 environments are taken into account while aggregating records. An Intel 5300 Network Interface Card (NIC) is always used to capture CSI, the most used RF signal for AR[84].

Table IV.isa summary of the traits, benefits, and drawbacks of several RF communication methods.



Technology	Device	Data	Description	Merits	Demerits
RFID	- Mobile device	CSI, Phase, RSS, TDoA	It stores and retrieves data via the electromagnetic broadcast to an RF - consistent, cohesive circuit	- High accuracy - Low cost - Power-efficient/free	- Tedious deployment - Short distances - Portable devices
Radar	Doppler radar	Doppler effect	It broadcast single-tone RF signals without involving modulation.	- simple design - power efficient - easy to deploy - simple - penetrative	- Frequency shift extremely relies on the circular velocity - Range folding - High maintenance
	FMCW radar	Range and doppler information	It captures doppler and range information concurrently thereby appropriate for multi targets scenarios		- Limited range - Prone to interference from other signals - Signal attenuation
	Interferometry radar	Micro Doppler signatures	It captures angular velocity using an interferometric receiver consisting of two antennas with correlated output.		- increased noise
	UWB radar	RF pulses	It broadcast an Rf signal with 25% greater fractional bandwidth.	- fine range resolution - extricate the target's scattering midpoints - penetrative - low electromagnetic radiation - power efficient	- Higher cost - Relative complexity - Special equipment - Hardly popularized
WiFi	- Routers - Access point - Mobile device	CSI	- Comprise amplitude and phase sub-signals represent the signal echoes of the human in subcarrier degree	- Wider range - Low cost - Comfortable - privacy-preserving - CSI high granularity - Easy to implement	- High false alarm ratio - RSS coarse granularity - RSS limited performance - Sensitive to slight changes in the environment
		RSS	- Change in the received signal strength in the receiver		

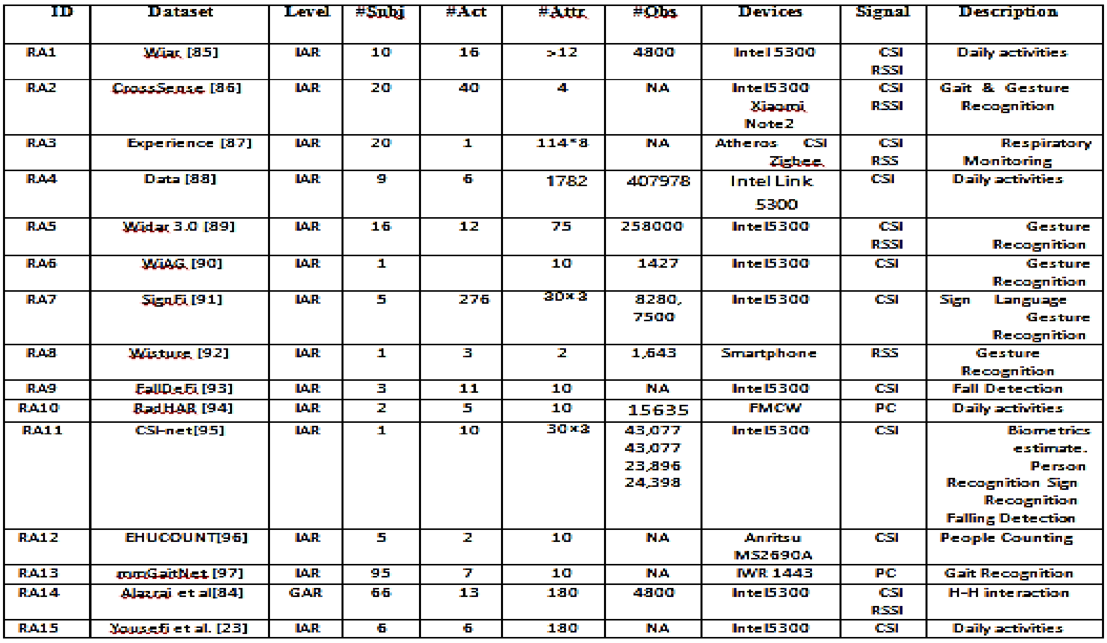
4.4 Indoor Positioning and tracking datasets

Similar to this, each of the three categories—sensory data ,vision data, or RF data—is represented by one of the publicly available datasets for the indoor location. Because RF data (CSI and RSS) is so effective at simulating human positioning information contrasted with data from sensors or vision, it is obvious that it dominates the current benchmarks. There aren't many indoor positioning benchmarks, and the majority of them are historical data, in contrast to activity recognition benchmarks.

5. Indoor IoT applications

Smart indoors makes it feasible to connect the numerous ubiquitous IoT devices included in many interior products, like smartphones, smart TVs, and smart refrigerators. Due to recent advancements in DL, researchers are employing it to solve a variety of smart indoor problems that contribute to improving quality of life through a variety of applications of smart indoor settings. This section offers numerous deep-learning approaches for various classes of user-centric IoT applications in smart indoor environments.

Table V. A survey of RF-based datasets for identifying indoor activities



ID	Dataset	Level	#Subj	#Act	#Attr	#Obs	Devices	Signal	Description
RA1	Wiar [85]	IAR	10	16	>12	4800	Intel 5300	CSI RSSI	Daily activities
RA2	CrossSense [86]	IAR	20	40	4	NA	Intel5300 Xiaomi Note2	CSI RSSI	Gait & Gesture Recognition
RA3	Experience [87]	IAR	20	1	114*8	NA	Atheros CSI Zigbee	CSI RSS	Respiratory Monitoring
RA4	Data [88]	IAR	9	6	1782	407978	Intel Link 5300	CSI	Daily activities
RA5	Widar 3.0 [89]	IAR	16	12	75	258000	Intel5300	CSI RSSI	Gesture Recognition
RA6	WiAG [90]	IAR	1		10	1427	Intel5300	CSI	Gesture Recognition
RA7	SignFi [91]	IAR	5	276	30×3	8280, 7500	Intel5300	CSI	Sign Language Gesture Recognition
RA8	Wisture [92]	IAR	1	3	2	1,643	Smartphone	RSS	Gesture Recognition
RA9	FallDeFi [93]	IAR	3	11	10	NA	Intel5300	CSI	Fall Detection
RA10	RadHAR [94]	IAR	2	5	10	15635	FMCW	PC	Daily activities
RA11	CSI-net[95]	IAR	1	10	30×3	43,077 43,077 23,896 24,398	Intel5300	CSI	Biometrics estimate. Person Recognition Sign Recognition Falling Detection
RA12	EHUCOUNT[96]	IAR	5	2	10	NA	Anritsu MS2690A	CSI	People Counting
RA13	mmGaitNet [97]	IAR	95	7	10	NA	IWR 1443	PC	Gait Recognition
RA14	Alazrai et al[84]	GAR	66	13	180	4800	Intel5300	CSI RSSI	H-H interaction
RA15	Yousefi et al. [23]	IAR	6	6	180	NA	Intel5300	CSI	Daily activities

H-H=” human-human” , H-O=” human-object” , NA=” not exist”

5.1 Vision-based Indoor Positioning

For a wide range of applications, CV methods' dependability and durability are widely known. Zhao et al. [107] investigate the prospect of enhancing localization performance by merging camera data with smartphone data and WiFi data to create a multimodal framework. This framework can help compute the interior vision of buildings for later localization or navigation reconstruction. Building Information Modeling (BIM) and CNNs are used to build a benchmark of compressed BIM photographs and Haite. [108] offer a unique visual indoor localization framework that analyzes the data to identify the nearest indoor photography equivalents. This is an approximation of the photo's orientation and indoor location. Another investigation by Chhikara et al. Unmanned aerial vehicle (UAV) indoor localization utilizing a CNN, Using a design for transfer learning, and employing a genetic algorithm to tune the model's hyperparameters. However, determining local coordinates and distances is not best done using visual data. Additionally, it is well known that visual data violates privacy and is restricted by LOS. Therefore, vision-based navigation, steering, and positioning techniques become less desirable option.

5.1.1 Sensor-based Indoor Positioning

Researchers take into account sensory data collected by numerous sensors installed in smartphones, smartwatches, and other devices to address the shortcomings of vision-based indoor localization techniques. As an illustration, A DL framework for localization utilizing magnetometer-gathered geomagnetic data was proposed by the authors of [110]. Following the encoding of the data into a repeating graph form, a CNN was used to automatically extract features and then classify the data. A deep learning system that learns how to perform indoor localization based on bimodal sensory data, including data acquired from light sensors and magnetometers, is described in [111]. This system uses LSTMs. This data can be used to enhance localization performance, according to experimental analyses of private datasets. To examine other sensor modalities, the authors of [112] proposed a multi-sensor DL framework that incorporates learning from multi-modal data from magnetometers, barometric pressure sensors, pressure sensors, and the Global Navigation Satellite System (GNSS). The framework makes use of MLPs to compute classification judgments, LSTMs to describe temporal dependencies, and dense convolutions for effective feature extraction. Pedestrian dead reckoning (PDR) has recently been acknowledged as one of the common ways to achieve indoor localization due to the extensive use of smart devices. The SAE network of [25] uses smartphone data (acceleration and gyroscope data) to estimate stride length in a PDR system in this regard.

In summary, the selection of DL models and sensors for localization is significantly influenced by a number of factors, such as efficiency, the environment for localization, the availability of computational resources, latency difficulties, etc. For instance, the usage of acceleration can be used to accomplish effective positioning. Magnetometer and gyroscope sequences. Wearing the sensor constantly, however, can be highly challenging for many. Knowing the ground plane can help with localization or tracking because individuals must move between levels in a multi-story workplace. A decision on the suitable DL layer, such as H. Convolutional layers, repetition layers, and attention layers, is also necessary given the significance of spatial information or temporal dependencies in the input.

5.1.2 RF-based indoor positioning and Tracking.

Triangulation and fingerprinting are the two subcategories of RF-based indoor localization that can be distinguished [12]. The fingerprinting process uses both online and offline CSI data. The measured CSI measurements are compared with the fingerprint data to determine the location of the target during the online phase or for tracking reasons when the system computes CSI reports of target locations throughout the offline phase to establish a fingerprint benchmark. On the other hand, triangulation/geometric approaches use the geometric features of triangles to determine and track a person's position. Table 7 shows that DL models based on RF signals have become more popular than those based on visual or sensory input.

5.2 Activity recognition

To effectively represent human-computer interaction in intelligent human-centric systems for surveillance, disability care, and healthcare, activity recognition (AR) is receiving growing scientific interest. The system, etc., has several advantages. IoT devices, smart technology, and augmented reality (AR) are now crucial components of efforts to enhance human existence in intelligent interior environments. Human activity, as we all know, is a conscious, aware, and personally significant series of actions that can be carried out by linked or unrelated people or groups. According to the complexity of the activities, the task of AR can be separated into three primary levels: individual activity detection (IAR), group activity detection (GAR), and hybrid activity detection (HAR). In IAR, the lowest degree of sophistication in AR, the main objective is to detect and identify actions taken by a single agent, regardless of actions taken by other agents in the environment. Additionally, in GAR, two or more individuals are actively involved in a variety of activities that may or may not be connected. Activities that are shared by subjects in related groups are those that help them all work toward a common objective. For instance, cooperation is required when several individuals are lifting a big object from the floor to a small table. Contrarily, independent group activity denotes some of the subject's behaviors as being autonomous and separate from those of others.For instance, each person's activities are independent of one another while they study, rest, watch TV, etc. [126]. Additionally, HAR's primary objective is to find indoor individual and group activities, which is a more difficult assignment. For instance, a smart home has three occupants, two of whom are sleeping and one of whom is cleaning dishes in the kitchen. As a result, a variety of both solo and group activities happen here. To this purpose, as was previously said, the research community has chosen to make use of deep learning to create effective and entirely automatic approaches to recognize various human activities.

5.2.1Vision-based activity recognition

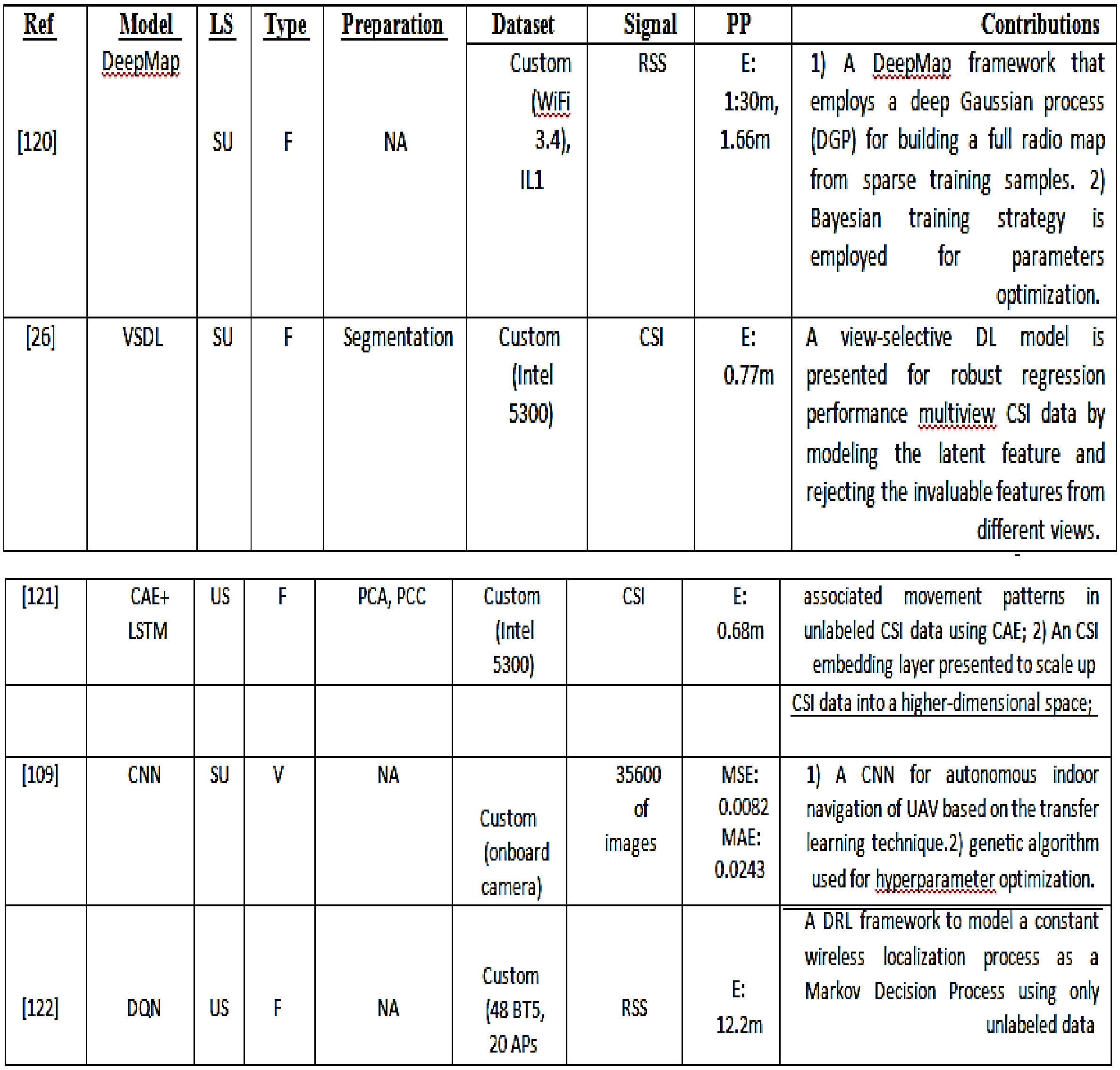
Vision-based deep learning methods mainly rely on visual sensor tools to monitor and record different types of indoor human activities [2]. What is most striking about this approach is that it is highly dependent on the quality of the captured image or recorded video. In simple terms, the primary elements that affect the quality of visual data are resolution, lighting circumstances, illumination variations, and comparable graphical features. As a result, computer vision researchers are more motivated than ever to develop a fresh approach to enhance AR performance from visual data with a manageable processing workload to satisfy the requirements of the IoT context. In terms of DL models, learning techniques (LS), AR layers, data preparation (preprocessing and/or feature engineering), and dataset IDs, Table 8 covers recent DL work on activity detection.

5.2.2 Sensor-based activity recognition

Recent developments and the widespread use of sensor technology have made sensor-based AR a more alluring research topic for the DL community. Since sensor data is smaller, processing it takes less time. Table 9 lists current developments in the DL modeling of several human activities from sensory input. This contains data preparation, dataset identification, reported accuracy, contributions, DL models, LS, and AR layers.

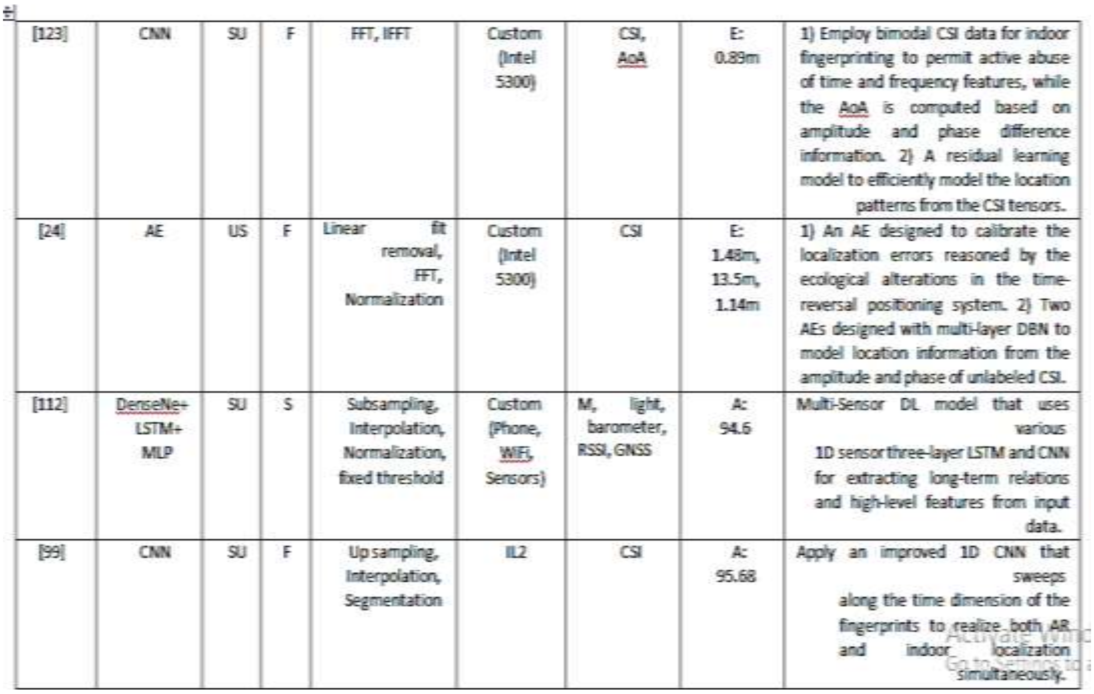
It is important to note that the majority of the current studies primarily train their models using supervised methods. Some of them are thinking about learning from readily accessible, enormous amounts of unlabeled sensory input. Through semi-supervised training, which uses both a significant amount of unannotated data and a small amount of labeled data during training, the authors of [137], [138], and [139] seek to overcome this issue. As an alternative, The hierarchical k-Medoids clustering method (Hk-mC) was used by the authors of [140] to produce an online learning LSTM model for unsupervised training on samples without tags. This approach automatically labels raw signals and generates hierarchical classifications. However, there hasn't been enough research done to fully understand the potential of unsupervised/semi-supervised DL for AR. The great majority of the papers we examined also concentrated on IAR-level AR. In an exceptional attempt to solve GAR, the authors of [141] used temporal convolutional networks (TCNs) and LSTM networks to predict multi-user behavior from tailored 2D light detection and ranging (LiDAR) data.

Table VII. Summary of indoor positioning and tracking investigations using deep learning



Ref	Model	LS	Type	Preparation	Dataset	Signal	PP	Contributions
[120]	DeepMap	SU	F	NA	Custom (WiFi 3.4), IL1	RSS	E: 1:30m, 1.66m	1) A DeepMap framework that employs a deep Gaussian process (DGP) for building a full radio map from sparse training samples. 2) Bayesian training strategy is employed for parameters optimization.
[26]	VSDL	SU	F	Segmentation	Custom (Intel 5300)	CSI	E: 0.77m	A view-selective DL model is presented for robust regression performance multiview CSI data by modeling the latent feature and rejecting the invaluable features from different views.
[121]	CAE+ LSTM	US	F	PCA, PCC	Custom (Intel 5300)	CSI	E: 0.68m	associated movement patterns in unlabeled CSI data using CAE; 2) An CSI embedding layer presented to scale up CSI data into a higher-dimensional space;
[109]	CNN	SU	V	NA	Custom (onboard camera)	35600 of images	MSE: 0.0082 MAE: 0.0243	1) A CNN for autonomous indoor navigation of UAV based on the transfer learning technique.2) genetic algorithm used for hyperparameter optimization.
[122]	DQN	US	F	NA	Custom (48 BT5, 20 APs	RSS	E: 12.2m	A DRL framework to model a constant wireless localization process as a Markov Decision Process using only unlabeled data

LS=" Learning Strategy", SU=" Supervised", US=" Unsupervised", SS=" Semi-supervised", PP= "positioning performance", A=" Accuracy", E=" Localization error", F=" fingerprinting", V=" Vision", S=" Sensor", T=" Triangulation", NA=" not exist"



[123]	CNN	SU	F	FFT, IFFT	Custom (Intel 5300)	CSI, AoA	E: 0.89m	1) Employ bimodal CSI data for indoor fingerprinting to permit active abuse of time and frequency features, while the AoA is computed based on amplitude and phase difference information. 2) A residual learning model to efficiently model the location patterns from the CSI tensors.
[24]	AE	US	F	Linear fit removal, FFT, Normalization	Custom (Intel 5300)	CSI	E: 1.48m, 13.5m, 1.14m	1) An AE designed to calibrate the localization errors reasoned by the ecological alterations in the time-reversal positioning system. 2) Two AEs designed with multi-layer DBN to model location information from the amplitude and phase of unlabeled CSI.
[112]	DenseNe+ LSTM+ MLP	SU	S	Subsampling, Interpolation, Normalization, fixed threshold	Custom (Phone, WiFi, Sensors)	M, light, barometer, RSSI, GNSS	A: 94.6	Multi-Sensor DL model that uses various 1D sensor three-layer LSTM and CNN for extracting long-term relations and high-level features from input data.
[99]	CNN	SU	F	Up sampling, Interpolation, Segmentation	IL2	CSI	A: 95.68	Apply an improved 1D CNN that sweeps along the time dimension of the fingerprints to realize both AR and indoor localization simultaneously.

[25]	SAE	SU	S	Segmentation, Interpolation	Custom (phone)	A, G data	E: 3.01	Deep AE for estimating step length by considering various walking velocities, the way the phone is carried, and the subject features.
[124]	ResNet+ LSTM	SU	F	Min-max normalization	IL3	RSSI	E: 3.20m	A spatial-temporal DL to learn both the spatial and temporal features using residual CNN and LSTM, respectively
[125]	CNN	SU	S	Sensor calibration Coordinate transformation	Custom (phone)	A, G, M data	1.06 m	A multi-head CNN is presented to extract walking patterns from input sequences, while the attention layer is employed to learn the relevance of convolutional features.
[114]	FFNN+ Fuzzy	SU	F	NA	Custom	RSS	MSE:3.20 MAE:1.36	A deep fuzzy forest model is presented to integrate the decision trees with FFNN to empower the representation learning capability.
[113]	CNN	SU	F	Phase calibration, Imaging	Custom (Intel 5300)	CSI (AoA)	E: 1.78m, 2.38m	Employ a CNN for indoor localization from imaged AoA values extracted from the phase of CSI data.
[115]	LSTM	SU	T	Normalization	DecaWave DW1000	UWB (TDoA)	AUC 0.997	A DL framework to handle the TDOA incorrect or missed measurements during asynchronous localization is called DeepTAL.
[116]	DNN	SU	T	Kalman filter, Distribution Judging, removing the invalids	Custom	RSS +TDOA	RMSE: 0.98	1) An enhanced RSS extraction technique to get more steady RSS values. 2) TDOA-based rapid discovery Procedure to calculate a coarse estimation of the target location.

5.2.3 RF-based activity recognition

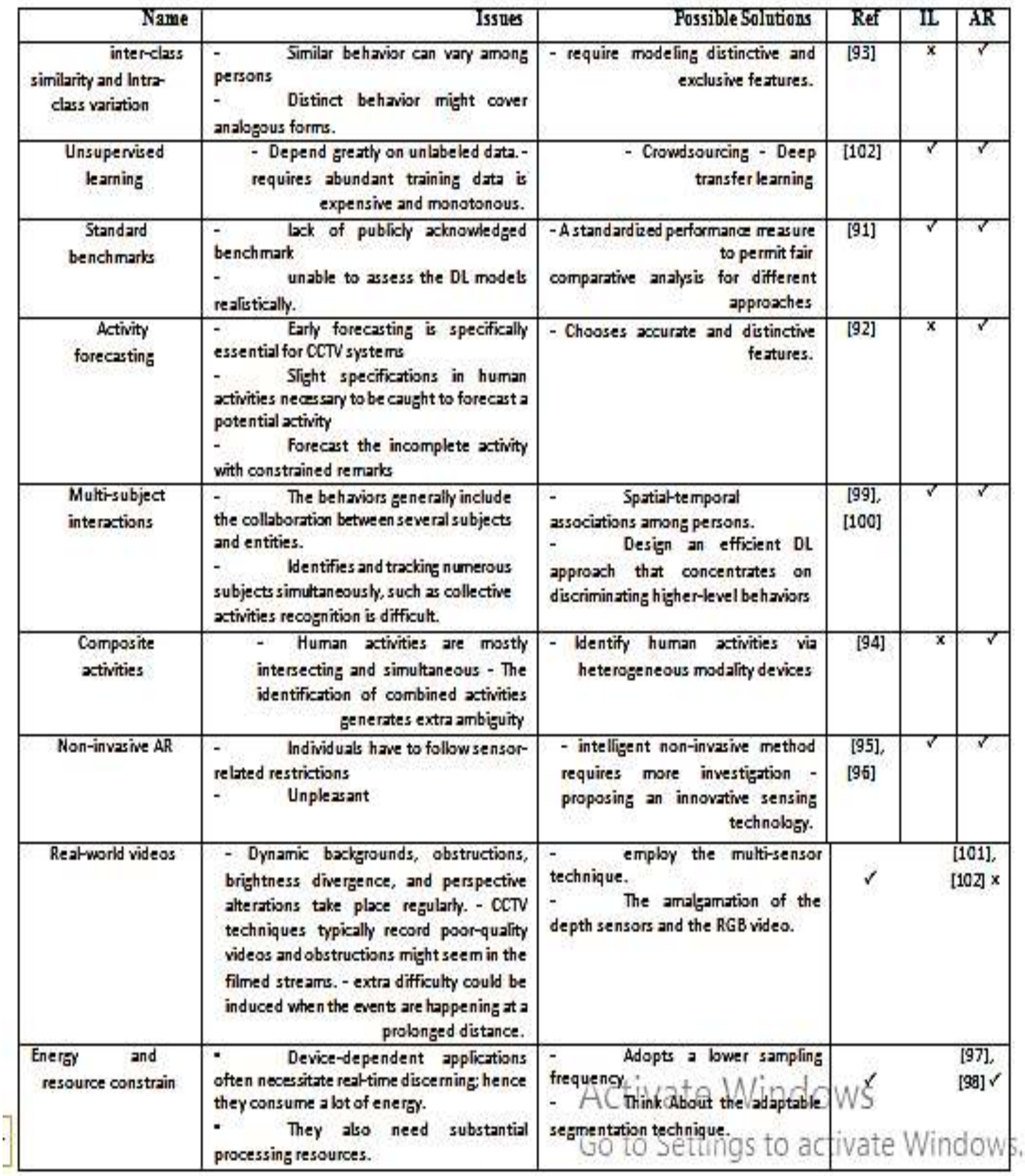
The use of radio waves to record human behavior has specific advantages due to its wide availability, which allays privacy concerns raised by vision-based and sensor-based techniques. Because walls and darkness can also interfere with RF signals, they are perfect for replicating human activity, especially inside. In recent years, RF-based augmented reality has gained popularity as a research topic. Commercial radio frequency solutions for sensing, capturing, and detecting many forms of human activity have been presented. Recent DL investigations on AR from various RF frequencies are compiled in Table 10.

It is clear that all of the DL studies under consideration place a strong emphasis on identifying individual user activities (IAR level), whereas GAR and HAR do not investigate areas. This might be a result of how difficult it is to record the varying multi-user activity in wireless signals. HF-based techniques employ supervised training like sensor-based techniques, while GANs created for fall and gesture recognition use semi-supervised training [153]. It is possible to enhance learning from unlabeled RF data by looking more closely at semi-supervised and unsupervised models. Additionally, the majority of studies examined experiments and validated their models using their unique datasets; however, it is vital to replicate their findings using open data to comprehend the advantages and disadvantages of their models. Additionally, CSI is well known for detecting human activity among multiple RF data types, the majority of which are recorded by the Intel 5300 NIC.

6. Emerging matters and future directions

The most fascinating research fields for indoor Internet of Things applications, encompassing both device-specific and generic applications approaches, are highlighted in this section. The primary challenges in developing intelligent indoor IoT applications are listed in Table XI, along with based on current research on intelligent IoT, potential solutions.

Table XI: lists difficulties and potential remedies for various IoT applications in indoor settings.



Name	Issues	Possible Solutions	Ref	IL	AR
inter-class similarity and Intra-class variation	- Similar behavior can vary among persons - Distinct behavior might cover analogous forms.	- require modeling distinctive and exclusive features.	[93]	×	✓
Unsupervised learning	- Depend greatly on unlabeled data.- requires abundant training data is expensive and monotonous.	- Crowdsourcing - Deep transfer learning	[102]	✓	✓
Standard benchmarks	- lack of publicly acknowledged benchmark - unable to assess the DL models realistically.	- A standardized performance measure to permit fair comparative analysis for different approaches	[91]	✓	✓
Activity forecasting	- Early forecasting is specifically essential for CCTV systems - Slight specifications in human activities necessary to be caught to forecast a potential activity - Forecast the incomplete activity with constrained remarks	- Chooses accurate and distinctive features.	[92]	×	✓
Multi-subject interactions	- The behaviors generally include the collaboration between several subjects and entities. - Identifies and tracking numerous subjects simultaneously, such as collective activities recognition is difficult.	- Spatial-temporal associations among persons. - Design an efficient DL approach that concentrates on discriminating higher-level behaviors	[99], [100]	✓	✓
Composite activities	- Human activities are mostly intersecting and simultaneous - The identification of combined activities generates extra ambiguity	- Identify human activities via heterogeneous modality devices	[94]	×	✓
Non-invasive AR	- Individuals have to follow sensor-related restrictions - Unpleasant	- intelligent non-invasive method requires more investigation - proposing an innovative sensing technology.	[95], [96]	✓	✓
Real-world videos	- Dynamic backgrounds, obstructions, brightness divergence, and perspective alterations take place regularly. - CCTV techniques typically record poor-quality videos and obstructions might seem in the filmed streams. - extra difficulty could be induced when the events are happening at a prolonged distance.	- employ the multi-sensor technique. - The amalgamation of the depth sensors and the RGB video.	✓		[101], [102] ×
Energy and resource constrain	▪ Device-dependent applications often necessitate real-time discerning; hence they consume a lot of energy. ▪ They also need substantial processing resources.	- Adopts a lower sampling frequency. - Think About the adaptable segmentation technique.	✓		[97], [98] ✓

6.1 Transfer learning

The development of intelligent IoT as a current trend in the computer vision community has been dominated by DL techniques. However, learning new DL methods from the start is still a difficult effort to create trustworthy applications. Implementing DL techniques that rely on prior pre-trained architectures that have already encountered the underlying data format is, therefore, a credible tactic. When employing visual or sensory data streams for indoor trending applications, it might be interesting to investigate transfer learning ideas.

6.2 Explainable deep learning

In recent years, research on the interpretability of visual models has become of utmost importance. There aren't many studies on interpretable video recognition methods, though. The detection of indoor activities, gestures, or indoor locations in a sequence of video frames derived from the target movie only requires a small number of keyframes, as described in [85, 86]. Furthermore, the related temporal characteristics of indoor activities and gestures vary. Based on the frames recorded at the start or conclusion of the video, it is possible to identify some action or gesture. Expert research can address issues like B. Frame arrangement in the time domain by examining the interpretability of complex activities/gestures based on keyframes. What impact do keyframes have on categorization choices? Is it possible to tailor these frameworks for quick teaching of DL techniques without sacrificing the effectiveness of indoor applications? Researchers can use this knowledge to create indoor IoT applications that are more effective.

6.3 Multimodal data

Multimodal data, such as the typical audio, image, text, and signaling data created and received by humans to interact with their surroundings, is frequently found in indoor locations. Reading, for instance, enables the rebuilding of a consistent component of a person's visual intelligence. As multimodal data provide intriguing semantic information, it is desirable to use multimodal information to comprehend complex interior activities [87]. Since learning directly from multimodal data can be challenging, modeling this type of data enables the collectionof long-term temporal connections between entities from multimodal data [88]. This long-term temporal dependence can show how indoor actions, gestures, and locations are sequentially ordered over time, much like how the human brain functions. One element from the lengthy main sequence initiates the following element once something is remembered, much like a persistent video. Furthermore, comprehending temporal interdependencies requires an understanding of interactions between various entities. For instance, a predefined interaction between items happens in a certain activity under a certain set of circumstances. Therefore, human multimodal information should be taken into account by DL-based indoor IoT applications to achieve dependable performance, particularly in applications that depend on long-term data.

6.4 The physical aspect of humans

The study of the physical components of human behavior, such as distinct and intricate movements and gestures, is becoming more and more popular today. For instance, the authors of [89] suggested a HAR dataset of 20 billion things to encourage researchers to investigate the link between humans and items. The record contains a class schema or a description of a document, like B. "Put an entity near an object" refers to a situation in which two things or a person interact. These data aid in the creation of indoor Internet of Things (IoT) applications that consider the physical characteristics of human movement and action as well as interactions with objects and spatial relationships. Closed-circuit television (CCTV) video may record a lot of statistics, but it can be challenging to record some physical characteristics including strength, speed, movement patterns, and speed. Therefore, it is crucial to create IoT benchmarks that take this data into account.

6.5 Learning actions without labels

Physically annotating data samples to expand the number of indoor datasets for training DL models from arbitrary application domains is time-consuming, ineffective, and expensive. Even while some regions allow for automatic annotation using search engines and video subtitles, human approval is still necessary. Crowdsourcing [104] is thought to be a more beneficial solution. However, label multiplicity concerns make it difficult and produce irrational outcomes. As a result, the research community must use more advanced and potent learning approaches that inevitably change created indoor data that is not labeled [140].

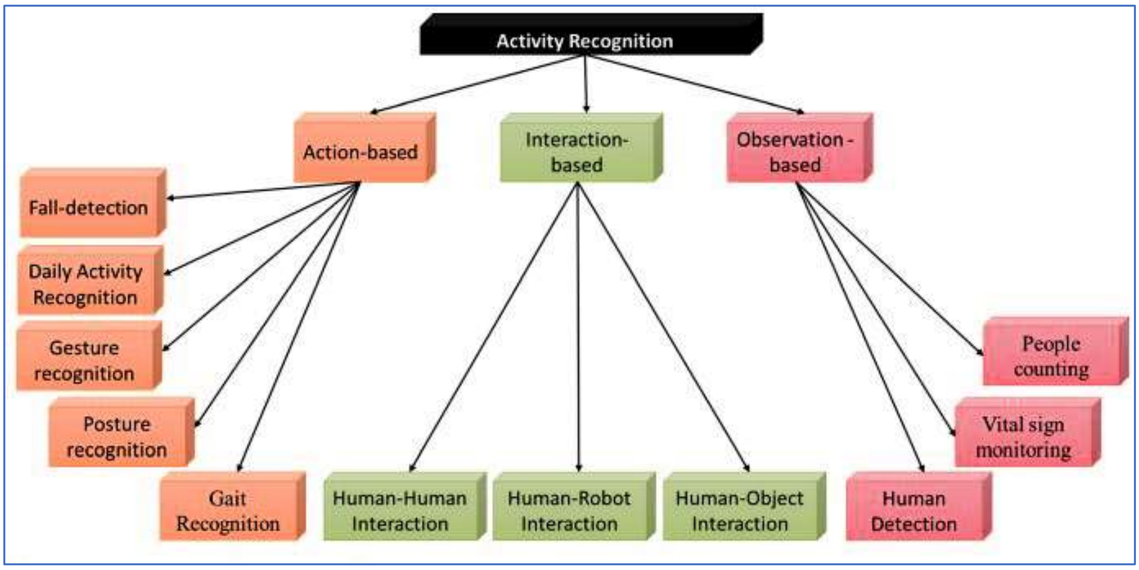


Figure (1) Human-centric activity recognition taxonomy in intelligent indoor environments

7. Conclusion

This article,may give a comprehensive overview of the most advanced deep learning algorithms and list the advantages and disadvantages of each from both device-dependent and device-independent viewpoints. Because they might be used in a range of IoT applications in smart indoor settings, such as geolocation and activity detection applications, these techniques have gained considerable attention in recent years. In the era of human-centric IoT applications in indoor environments, comprehensive explanations, analyses, and insights into pertinent features aid academics in expanding their knowledge.

When addressing current research, a number of factors are taken into account, such as deep learning construction, precision, programs, settings, data used, sensors, and examples. We value the most recent developments in IoT applications that involve and don't involve specific devices. In terms of data modalities and/or application domains, we present a new taxonomy for intelligent indoor DL approaches. Investigated are the features, benefits, and drawbacks of contemporary deep learning techniques applied to indoor IoT applications. The most fascinating research questions in indoor IoT applications are also examined in this review study, and viable solutions are offered.

A number of difficult subjects, including B. System Design, Tracking of Multiple Activities, Motion Prediction, and Temporal Sensitivity, are beneficial for future research in addition to deep learning applications in varied indoor situations. Research on numerous human-centric indoor IoT applications may be stimulated by this study.

References

[1] Abuhamad, M., Abusnaina, A., Nyang, D., &Mohaisen, D. (2021). Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 30200 76

[2] Alam, F., Faulkner, N., & Parr, B. (2021). Device-free localization: A review of non-RF techniques for unobtrusive indoor positioning. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020.30301 74

[3] Alazrai, R., Awad, A., Alsaify, B., Hababeh, M., &Daoud, M. I. (2020). A dataset for Wi-Fi-based human-to-human interaction recognition. Data in Brief. https:// doi. org/ 10. 1016/j. dib. 2020. 105668

[4] Alemdar, H., Ertan, H., Incel, O. D., &Ersoy, C. (2013). ARAS human activity datasets in multiple homes with multiple residents. https:// doi. org/ 10. 4108/ icst. pervasivehealth. 2013. 252120.

[5] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones.

[6] Bai, J., Lian, S., Liu, Z., Wang, K., & Liu, D. (2017). Smart guiding glasses for visually impaired people in indoor environment. IEEE Transactions on Consumer Electronics. https:// doi. org/ 10. 1109/ TCE. 2017.014980

[7] Banos, O., et al. (2015). Design, implementation, and validation of a novel open framework for agile development of mobile health applications. Biomedical Engineering Online. https:// doi. org/ 10. 1186/1475- 925X- 14- S2- S6

[8] Barshan, B., &Yüksek, M. C. (2013). Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. The Computer Journal. https:// doi. org/ 10. 1093/comjnl/ bxt075

[9] Barut, O., Zhou, L., &Luo, Y. (2020). Multi-task LSTM model for human activity recognition and intensity estimation using wearable sensor data. IEEE Internet of Things Journal,7(9), 8760–8768. https:// doi.org/ 10. 1109/ JIOT. 2020. 29965 78

[10] Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N., &Raffel, C. (2019). MixMatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems, 32.

[11] Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M., & De Munari, I. (2019). IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2019. 29202 83

[12] Blunsden, S., & Fisher, R. B. (2010). The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. Annals of the BMVA, 4, 1–12.

[13] Brinke, J. K., &Meratnia, N. (2019). Dataset: Channel state information for different activities, participants and days. https:// doi. org/ 10. 1145/ 33594 27. 33619 13.

[14] Carreira, J., Noland, E., Hillier, C., &Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. July 2019, [Online]. http:// arxiv. org/ abs/ 1907. 06987.

[15] CASAS Smart Home Project. (2021). http:// casas. wsu. edu/ datasets/. Accessed March 25, 2021.

[16] Chavarriaga, R., et al. (2013). The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. Pattern Recognition Letters. https:// doi. org/ 10. 1016/j. patrec. 2012. 12. 014

[17] Chen, C., Jafari, R., &Kehtarnavaz, N. (2015). UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. https:// doi. org/ 10. 1109/ ICIP. 2015. 73507 81.

[18] Chen, D., Yongchareon, S., Lai, E. M. K., Yu, J., & Sheng, Q. Z. (2021). Hybrid fuzzy C-means CPD-based segmentation for improving sensor-based multi-resident activity recognition. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2021. 30515 74

[19] Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., &Nie, F. (2020). A semisupervised recurrent convolutional attention model for human activity recognition. IEEE Transactions on Neural Networks and Learning Systems. https:// doi. org/ 10. 1109/ TNNLS. 2019. 29272 24

[20] Chen, M., et al. (2020). MoLoc: Unsupervised fingerprint roaming for device-free indoor localization in a mobile ship environment. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 30042 40

[21] Chen, Z., Zhang, L., Jiang, C., Cao, Z., & Cui, W. (2019). WiFi CSI Based passive human activity recognition using attention based BLSTM. IEEE Transactions on Mobile Computing. https:// doi. org/ 10.

[22] 1109/ TMC. 2018. 28782 33

[23] Cheplygina, V., de Bruijne, M., &Pluim, J. P. W. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis. https:// doi. org/ 10. 1016/j. media. 2019. 03. 009

[24] Chhikara, P., Tekchandani, R., Kumar, N., Chamola, V., &Guizani, M. (2021). DCNN-GA: A deep neural net architecture for navigation of UAV in indoor environment. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 30270 95

[25] Choi, W., Shahid, K., &Savarese, S. (2009). What are they doing? Collective activity classification using spatio-temporal relationship among people. https:// doi. org/ 10. 1109/ ICCVW. 2009. 54574 61.

[26] Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and visionbased human activity recognition: A comprehensive survey. Pattern Recognition. https:// doi. org/ 10. 1016/j. patcog. 2020. 107561

[27] Deep, S., Zheng, X., Karmakar, C., Yu, D., Hamey, L. G. C., & Jin, J. (2020). A survey on anomalous behavior detection for elderly care using dense-sensing networks. IEEE Communications Surveys & Tutorials. https:// doi. org/ 10. 1109/ COMST. 2019. 29482 04

[28] Dhall, A., Goecke, R., &Gedeon, T. (2015). Automatic group happiness intensity analysis. IEEE Transactions on Affective Computing. https:// doi. org/ 10. 1109/ TAFFC. 2015. 23974 56

[29] Dhiman, C., &Vishwakarma, D. K. (2020). View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Transactions on Image Processing. https:// doi. org/ 10. 1109/ TIP. 2020. 29652 99

[30] Feng, C., Arshad, S., Zhou, S., Cao, D., & Liu, Y. (2019). Wi-Multi: A three-phase system for multiple human activity recognition with commercial WiFi devices. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2019. 29159 89

[31] Gao, N. et al. (2020).Generative adversarial networks for spatio-temporal data: A survey. arXiv. 2020.

[32] Gochoo, M., Tan, T. H., Liu, S. H., Jean, F. R., Alnajjar, F. S., & Huang, S. C. (2019). Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. EEE Journal of Biomedical and Health Informatics. https:// doi. org/ 10. 1109/ JBHI. 2018. 28336 18

[33] Gordon, D., Wirz, M., Roggen, D., Tröster, G., &Beigl, M. (2014). Group affiliation detection using model divergence for wearable devices. https:// doi. org/ 10. 1145/ 26343 17. 26343 19.

[34] Goyal, R. et al. (2017). The ‘something something’ video database for learning and evaluating visual common sense. https:// doi. org/ 10. 1109/ ICCV. 2017. 622.

[35] Gu, C., et al. (2018). AVA: A video dataset of spatio-temporally localized atomic visual actions. https:// doi.

[36] org/ 10. 1109/ CVPR. 2018. 00633.

[37] Gu, F., Khoshelham, K., Yu, C., & Shang, J. (2019). Accurate step length estimation for pedestrian dead reckoning localization using stacked autoencoders. IEEE Transactions on Instrumentation and Measurement. https:// doi. org/ 10. 1109/ TIM. 2018. 28718 08

[38] Guo, L., et al. (2019). Wiar: A public dataset for wifi-based activity recognition. IEEE Access. https:// doi. org/ 10. 1109/ ACCESS. 2019. 29470 24

[39] Guo, T., Xu, C., He, S., Shi, B., Xu, C., & Tao, D. (2020b). Robust student network learning. IEEE Transactions on Neural Networks and Learning Systems. https:// doi. org/ 10. 1109/ TNNLS. 2019. 29291 14

[40] Guo, X., Ansari, N., Hu, F., Shao, Y., Elikplim, N. R., & Li, L. (2020a). A survey on fusion-based indoor positioning. IEEE Communications Surveys & Tutorials. https:// doi. org/ 10. 1109/ COMST.

[41] 2019. 29510 36

[42] Guo, Z., Xiao, F., Sheng, B., Fei, H., & Yu, S. (2020c). WiReader: Adaptive air handwriting recognition based on commercial WiFi signal. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 29970 53

[43] Ha, I., Kim, H., Park, S., & Kim, H. (2018). Image retrieval using BIM and features from pretrained VGG network for indoor localization. Building and Environment. https:// doi. org/ 10. 1016/j. build env. 2018. 05. 026

[44] Haseeb, M. A. A., &Parasuraman, R. (2017). Wisture: RNN-based learning of wireless signals for gesture recognition in unmodified smartphones. arXiv. 2017.

[45] Hayashi, T., Nishida, M., Kitaoka, N., & Takeda, K. (2015). Daily activity recognition based on DNN using environmental sound and acceleration signals. https:// doi. org/ 10. 1109/ EUSIP CO. 2015. 73627 96.

[46] He, J., & So, H. C. (2020). A hybrid TDOA-fingerprinting-based localization system for LTE network.

[47] IEEE Sensors Journal. https:// doi. org/ 10. 1109/ JSEN. 2020. 30041 79

[48] He, Y., Chen, Y., Hu, Y., &Zeng, B. (2020). WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020.29894 26

[49] Heilbron, F. C., Escorcia, V., Ghanem, B., &Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. https:// doi. org/ 10. 1109/ CVPR. 2015. 72986 98.

[50] Hillyard, P., et al. (2018). Experience: Cross-technology radio respiratory monitoring performance study. https:// doi. org/ 10. 1145/ 32415 39. 32415 60.

[51] Huang, J., Lin, S., Wang, N., Dai, G., Xie, Y., & Zhou, J. (2020). TSE-CNN: A two-stage end-to-end CNN for human activity recognition. IEEE Journal of Biomedical and Health Informatics. https://doi. org/ 10. 1109/ JBHI. 2019. 29096 88

[52] Huang, Y., Kaufmann, M., Aksan, E., Black, M. J., Hilliges, O., & Pons-Moll, G. (2018). Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. https://doi. org/ 10. 1145/ 32721 27. 32751 08.

[53] Hussain, T., et al. (2020). Multi-view summarization and activity recognition meet edge computing in

[54] IoT environments. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ jiot. 2020. 30274 83

[55] Hussain, Z., Sheng, Q. Z., & Zhang, W. E. (2020). A review and categorization of techniques on devicefree human activity recognition. Journal of Network and Computer Applications. https:// doi. org/10. 1016/j. jnca. 2020. 102738

[56] Huynh-The, T., Hua, C. H., Tu, N. A., & Kim, D. S. (2021). Physical activity recognition with statisticaldeep fusion model using multiple sensory data for smart health. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 30132 72

[57] Jekabsons, G., &Zuravlyovs, V. (2010). Refining Wi-Fi based indoor positioning. In Aict2010—Application of Information and Communication Technologies Proceedings of 4Th International Science Conference, 2010.

[58] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. F. (2014). Large-scale video classification with convolutional neural networks. https:// doi. org/ 10. 1109/ CVPR. 2014. 223.

[59] Khan, P., Reddy, B. S. K., Pandey, A., Kumar, S., & Youssef, M. (2020). Differential channel-stateinformation-based human activity recognition in IoT networks. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2020. 29972 37

[60] Khan, A., Wang, S., & Zhu, Z. (2019). Angle-of-arrival estimation using an adaptive machine learning framework. IEEE Communications Letters. https:// doi. org/ 10. 1109/ LCOMM. 2018. 28844 64

[61] Kim, E. (2020). Interpretable and accurate convolutional neural networks for human activity recognition.

[62] IEEE Transactions on Industrial Informatics. https:// doi. org/ 10. 1109/ TII. 2020. 29726 28

[63] Kim, M., Han, D., & Rhee, J. K. (2021). Multiviewvariational deep learning with application to practical indoor localization. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2021. 30635 12

[64] Ko, W. R., Jang, M., Lee, J., & Kim, J. (2021). AIR-Act2Act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. The International Journal of Robotics Research. https:// doi. org/ 10. 1177/ 02783 64921 990671

[65] Koppula, H. S., Gupta, R., &Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. The International Journal of Robotics Research. https:// doi. org/ 10. 1177/ 0278364913 478446

[66] Kuehne, H., Jhuang, H., Stiefelhagen, R., &Serre Thomas, T. (2013). Hmdb51: A large video database for human motion recognition. In High performance computing in science and engineering' 12: Transactions of the high performance computing center, Stuttgart (HLRS) 2012.

[67] Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers.

[68] ACM SIGKDD Explorations Newsletter. https:// doi. org/ 10. 1145/ 19648 97. 19649 18

[69] Lee, N., Ahn, S., & Han, D. (2018). AMID: Accurate magnetic indoor localization using deep learning. Sensors (switzerland). https:// doi. org/ 10. 3390/ s1805 1598

[70] Leutheuser, H., Doelfel, S., Schuldhaus, D., Reinfelder, S., &Eskofier, B. M. (2014). Performance comparison of two step segmentation algorithms using different step activities. https:// doi. org/ 10. 1109/ BSN.2014. 37.

[71] Leutheuser, H., Schuldhaus, D., &Eskofier, B. M. (2013). Hierarchical, multi-sensor based classification of daily life activities: Comparison with state-of-the-art algorithms using a benchmark dataset. PLoS ONE. https:// doi. org/ 10. 1371/ journ al. pone. 00751 96

[72] Li, J., Xie, X., Pan, Q., Cao, Y., Zhao, Z., & Shi, G. (2020c). SGM-net: Skeleton-guided multimodal network for action recognition. Pattern Recognition. https:// doi. org/ 10. 1016/j. patcog. 2020. 107356

[73] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., &Tian, Q. (2021b). Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence. https:// doi. org/ 10. 1109/ TPAMI. 2021. 30537 65

[74] Li, Q., Gravina, R., Li, Y., Alsamhi, S. H., Sun, F., &Fortino, G. (2020a). Multi-user activity recognition: Challenges and opportunities. Information Fusion. https:// doi. org/ 10. 1016/j. inffus. 2020. 06. 004

[75] Li, X., Wang, Y., Zhang, B., & Ma, J. (2020d). PSDRNN: An efficient and effective HAR scheme based on feature extraction and deep learning. IEEE Transactions on Industrial Informatics. https:// doi. org/ 10.1109/ TII. 2020. 29689 20

[76] Li, X., Yu, L., Chen, H., Fu, C. W., Xing, L., &Heng, P. A. (2021a). Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems. https:// doi. org/ 10. 1109/ TNNLS. 2020. 29953 19

[77] Li, Y., Hu, X., Zhuang, Y., Gao, Z., Zhang, P., & El-Sheimy, N. (2020b). Deep reinforcement learning (DRL): Another perspective for unsupervised wireless localization. IEEE Internet of

Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2019. 29577 78

[78] Liu, J., Liu, H., Chen, Y., Wang, Y., & Wang, C. (2020a). Wireless sensing for human activity: A survey.

[79] IEEE Communications Surveys & Tutorials. https:// doi. org/ 10. 1109/ COMST. 2019. 29344 89

[80] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., &Kot, A. C. (2020b). NTU RGB+D 120: A largescale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and

[81] Machine Intelligence. https:// doi. org/ 10. 1109/ TPAMI. 2019. 29168 73

[82] Lohan, E. S., Torres-Sospedra, J., Leppäkoski, H., Richter, P., Peng, Z., & Huerta, J. (2017). Wi-Fi crowdsourced fingerprinting dataset for indoor positioning. Data. https:// doi. org/ 10. 3390/ data2 040032

[83] Lu, N., Wu, Y., Feng, L., & Song, J. (2019). Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data. IEEE Journal of Biomedical and Health Informatics. https:// doi. org/ 10. 1109/ JBHI. 2018. 28082 81

[84] Luo, F., Poslad, S., &Bodanese, E. (2020). Temporal convolutional networks for multiperson activity recognition using a 2-D LIDAR. IEEE Internet of Things Journal,7(8), 7432–7442. https:// doi. org/ 10.1109/ JIOT. 2020. 29845 44

[85] Ma, Y., Zhou, G., Wang, S., Zhao, H., & Jung, W. (2018). SignFi: Sign language recognition using WiFi. In Proceedings of ACM interactive, mobile, wearable ubiquitous technol, 2018. https:// doi. org/ 10. 1145/31917 55.

[86] Marszałek, M., Laptev, I., &Schmid, C. (2009). Actions in context. https:// doi. org/ 10. 1109/ CVPRW. 2009. 52065 57.

[87] Meng, F., Liu, H., Liang, Y., Tu, J., & Liu, M. (2019). Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition. IEEE Transactions on Image Processing. https:// doi. org/ 10. 1109/ TIP. 2019. 29135 44

[88] Meng, Z., et al. (2020). Gait recognition for co-existing multiple people using millimeter wave sensing (Vol.

[89] 34, No. 01, pp. 849–856). https:// ojs. aaai. org/ index. php/ AAAI/ artic le/ view/ 5430.

[90] Micucci, D., Mobilio, M., &Napoletano, P. (2017). UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones. Applied Sciences. https:// doi. org/ 10. 3390/ app71 01101 Mohammadi, M., Al-Fuqaha, A., Sorour, S., &Guizani, M. (2018).

Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys and Tutorials.https:// doi. org/ 10. 1109/COMST. 2018. 28443 41

[91] Monfort, M., et al. (2020). Moments in time dataset: One million videos for event understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence. https:// doi. org/ 10. 1109/ TPAMI. 2019.29014 64

[92] Montoliu, R., Sansano, E., Torres-Sospedra, J., & Belmonte, O. (2017). IndoorLoc platform: A public repository for comparing and evaluating indoor positioning systems. https:// doi. org/ 10. 1109/ IPIN.2017. 81159 40.

[93] Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., & Weber, A. (2007). Documentation mocap database hdm05, 2007.

[94] Nirmal, I., Khamis, A., Hassan, M., Hu, W., & Zhu, X. (2021). Deep learning for radio-based human sensing: Recent advances and future directions. IEEE Communications Surveys & Tutorials. https:// doi.org/ 10. 1109/ COMST. 2021. 30583 33

[95] Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Information Fusion. https:// doi. org/ 10. 1016/j. inffus. 2018. 06. 002

[96] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., &Bajcsy, R. (2013). Berkeley MHAD: A comprehensive multimodal human action database. https:// doi. org/ 10. 1109/ WACV. 2013. 64749 99.

[97] Oguntala, G., Hu, Y. F., Alabdullah, A. A. S., Abd-Alhameed, R., Ali, M., &Luong, D. (2021). Passive RFID module with LSTM recurrent neural network activity classification algorithm for ambient assisted living. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2021. 30512 47

[98] Palipana, S., Rojas, D., Agrawal, P., &Pesch, D. (2018). FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices. In Proceedings of ACM interactive, mobile, wearable ubiquitous technol, 2018. https:// doi. org/ 10. 1145/ 31611 83.

[99] Pei, L., et al. (2020). MARS: Mixed virtual and real wearable sensors for human activity recognition with multi-domain deep learning model. arXiv. 2020. https:// doi. org/ 10. 1109/ jiot. 2021. 30558 59.

[100] Qi, W., Su, H., &Aliverti, A. (2020). A smartphone-based adaptive recognition and real-time monitoring system for human activities. IEEE Transactions on Human-Machine. https:// doi. org/ 10. 1109/ THMS.2020. 29841 81

[101] Qian, K., Wu, C., Yang, Z., Liu, Y., & Jamieson, K. (2017). Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. https:// doi. org/ 10. 1145/ 30840 41. 30840 67.

[102] Qian, K., Wu, C., Zhang, Y., Zhang, G., Yang, Z., & Liu, Y. (2018). Widar2.0: Passive human tracking with a single Wi-Fi link. https:// doi. org/ 10. 1145/ 32102 40. 32103 14.

[103] Qin, Z., Zhang, Y., Meng, S., Qin, Z., &Choo, K. K. R. (2020). Imaging and fusing time series for wearable sensor-based human activity recognition. Information Fusion. https:// doi. org/ 10. 1016/j. inffus. 2019.06. 014

[104] Rashid, N., Dautta, M., Tseng, P., & Al Faruque, M. A. (2021). HEAR: Fog-enabled energy-aware online human eating activity recognition. IEEE Internet of Things Journal.https:// doi. org/ 10. 1109/ JIOT.2020. 30088 42.

[105] Reiss, A., &Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. https:// doi. org/ 10. 1109/ ISWC. 2012. 13.

[106] Rossi, S., Capasso, R., Acampora, G., & Staffa, M. (2018). A multimodal deep learning network for group activity recognition. https:// doi. org/ 10. 1109/ IJCNN. 2018. 84893 09.

[107] Ryoo, M. S., &Aggarwal, J. K. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. https:// doi. org/ 10. 1109/ ICCV. 2009. 54593 61.

[108] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. https:// doi. org/ 10. 1109/ CVPR. 2016. 115.

[109] Sheng, B., Fang, Y., Xiao, F., & Sun, L. (2020a). An accurate device-free action recognition system using two-stream network. IEEE Transactions on Vehicular Technology. https:// doi. org/ 10. 1109/ TVT. 2020.29939 01

[110] Sheng, B., Xiao, F., Sha, L., & Sun, L. (2020b). Deep spatial-temporal model based cross-scene action recognition using commodity WiFi. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT.2020. 29732 72

[111] Shu, X., Tang, J., Qi, G. J., Liu, W., & Yang, J. (2021a). Hierarchical long short-term concurrent memory for human interaction recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https:// doi. org/ 10. 1109/ TPAMI. 2019. 29420 30

[112] Shu, X., Zhang, L., Sun, Y., & Tang, J. (2021b). Host-parasite: graph LSTM-in-LSTM for group activity recognition. IEEE Transactions on Neural Networks and Learning Systems. https:// doi. org/ 10. 1109/TNNLS. 2020. 29789 42

[113] Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., &Alahari, K. (2018). Actor and observer: Joint modeling of first and third-person videos. https:// doi. org/ 10. 1109/ CVPR. 2018. 00772.

[114] Sikder, N., &Nahid, A.-A. (2021). KU-HAR: An open dataset for heterogeneous human activity recognition. Pattern Recognition Letters,146, 46–54. https:// doi. org/ 10. 1016/j. patrec. 2021. 02. 024

[115] Singh, A. D., Sandha, S. S., Garcia, L., &Srivastava, M. (2019). Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar.https:// doi. org/ 10. 1145/ 33496 24. 33567 68.

[116] Sobron, I., Del Ser, J., Eizmendi, I., & Velez, M. (2018). Device-free people counting in IoT environments: New insights, results, and open challenges. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/JIOT. 2018. 28069 90

[117] Sohn, I. (2021). Deep belief network based intrusion detection techniques: A survey. Expert Systems with Applications.https:// doi. org/ 10. 1016/j. eswa. 2020. 114170

[118] Sohn, K. et al. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. arXiv. 2020.

[119] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. December 2012, [Online]. http:// arxiv. org/ abs/ 1212. 0402.

[120] Stisen, A., et al. (2015). Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. https:// doi. org/ 10. 1145/ 28096 95. 28097 18.

[121] Sztyler, T.,&Stuckenschmidt, H. (2016) "On-body localization of wearable devices: An investigation of position-aware activity recognition. https:// doi. org/ 10. 1109/ PERCOM. 2016. 74565 21.

[122] Tang, J., Shu, X., Yan, R., & Zhang, L. (2019a). Coherence constrained graph LSTM for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https:// doi. org/ 10. 1109/tpami. 2019. 29285 40

[123] Tang, Y., Lu, J., Wang, Z., Yang, M., & Zhou, J. (2019b). Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE Transactions on Image Processing. https:// doi.org/ 10. 1109/ tip. 2019. 29145 77

[124] Tarvainen, A., &Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems, 30.

[125] Thariq Ahmed, H. F., Ahmad, H., &Cv, A. (2020). Device free human gesture recognition using Wi-Fi CSI: A survey. Engineering Applications of Artificial Intelligence. https:// doi. org/ 10. 1016/j. engappai. 2019.

[126] 103281

[127] Torres-Sospedra, J., et al. (2014). UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. https:// doi. org/ 10. 1109/ IPIN. 2014. 72754 92.

[128] Torres-Sospedra, J., et al. (2017). The smartphone-based offline indoor location competition at IPIN 2016: Analysis and future work. Sensors (Switzerland),10, 100. https:// doi. org/ 10. 3390/ s1703 0557

[129] Torres-Sospedra, J., Rambla, D., Montoliu, R., Belmonte, O., Huerta, J. (2015). UJIIndoorLoc-Mag: A new database for magnetic field-based localization problems. https:// doi. org/ 10. 1109/ IPIN. 2015. 73467 63.

[130] Uddin, M. Z., Hassan, M. M., Alsanad, A., &Savaglio, C. (2020). A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare. Inf. Fusion. https:// doi. org/ 10. 1016/j. inffus. 2019. 08. 004

[131] Ugulino, W., Cardador, D., Vega, K., Velloso, E., Milidiú, R., &Fuks, H. (2012). Wearable computing:

[132] Accelerometers’ data classification of body postures and movements. https:// doi. org/ 10. 1007/978-3- 642- 34459-6_6.

[133] Virmani, A. &Shahzad, M. (2017). Position and orientation agnostic gesture recognition using WiFi. https:// doi. org/ 10. 1145/ 30813 33. 30813 40.

[134] Wang, F., Feng, J., Zhao, Y., Zhang, X., Zhang, S., & Han, J. (2019a). Joint activity recognition and indoor localization with WiFi fingerprints. IEEE Access. https:// doi. org/ 10. 1109/ ACCESS. 2019. 29237 43

[135] Wang, F., Gong, W., & Liu, J. (2019c). On spatial diversity in wifi-based human activity recognition: A deep learning-based approach. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT. 2018.28714 45

[136] Wang, F., Han, J., Zhang, S., He, X., & Huang, D. (2018) “CSI-Net: Unified human body characterization and pose recognition. arXiv. 2018.

[137] Wang, F., Liu, J., & Gong, W. (2020e). Multi-adversarial in-car activity recognition using RFIDs. IEEE

[138] Transactions on Mobile Computing. https:// doi. org/ 10. 1109/ tmc. 2020. 29779 02

[139] Wang, Q., et al. (2021). Pedestrian dead reckoning based on walking pattern recognition and online magnetic fingerprint trajectory calibration. IEEE Internet of Things Journal. https:// doi. org/ 10. 1109/ JIOT.2020. 30161 46

[140] Wang, R. et al. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. https:// doi. org/ 10. 1145/ 26320 48. 26320 54.

[141] Wang, R., Luo, H., Wang, Q., Li, Z., Zhao, F., & Huang, J. (2020d). A spatial-temporal positioning algorithm using residual network and LSTM. IEEE Transactions on Instrumentation and Measurement. https:// doi. org/ 10. 1109/ TIM. 2020. 29986 45

[142] Wang, W., Bai, P., Zhou, Y., Liang, X., & Wang, Y. (2019b). Optimal configuration analysis of AOA localization and optimal heading angles generation method for UAV swarms. IEEE Access. https:// doi. org/10. 1109/ ACCESS. 2019. 29182 99