Deep Learning Approaches for Human-Centered IoT Applications in Smart Indoor Environments: A Contemporary Survey

Mohamed Abdel-Basset[1], Victor Chang[*2], Hossam Hawash[1], Ripon K. Chakrabortty [3] and Michael Ryan [3]
[1]Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt.
Emails: *mohamedbasset@zu.edu.eg; hossammoh@zu.edu.eg*
[2]School of Computing, Engineering and Digital Technologies, Teesside University, UK
Email: victorchang.research@gmail.com/V.Chang@tees.ac.uk
[3]Capability Systems Centre, School of Engineering and IT, UNSW Canberra; Australia
Emails: *r.chakrabortty@adfa.edu.au; m.ryan@adfa.edu.au*

Abstract
The widespread Internet of Things (IoT) technologies in day life indoor environments result in an enormous amount of daily generated data, which require reliable data analysis techniques to enable efficient exploitation of this data. The recent developments in deep learning (DL) have facilitated the processing and learning from the massive IoT data and learn essential features swiftly and professionally for a variety of IoT applications on smart indoor environments. This study surveys the recent literature on exploiting DL for different indoor IoT applications. We aim to give insights into how the DL approaches can be employed from various viewpoints to develop improved Indoor IoT applications in two distinct domains: indoor positioning/tracking and activity recognition. A primary target is to effortlessly amalgamate the two disciplines of IoT and DL, resultant in a broad range of innovative strategies in indoor IoT applications, such as health monitoring, smart home control, robotics, etc. Further, we have derived a thematic taxonomy from the comparative analysis of technical studies of the three beforementioned domains. Eventually, we proposed and discussed a set of matters, challenges, and some new directions in incorporating DL to improve the efficiency of indoor IoT applications, encouraging and stimulating additional advances in this auspicious research area**.**

Keywords: Deep learning; Internet of Things; Smart Indoor Environments, Activity Recognition, Indoor Positioning and Tracking.

I. INTRODUCTION

The continuous increase of population resided in urban areas where everyone has a busy, stressful day life that makes everyone looking forward to some comfortability and effortless life in their home and with the matter of fact that most of the people spend around 80% of their time indoors imposes many challenges on enhancing the day life quality of urban citizens. The idea of developing a smart building system is to exploit smart appliances to enhance the superiority of life quality of citizens in the indoor environment [1]. Accordingly, we have multiple smart home applications to improve residents' indoor life, such as controlling indoor appliances remotely, indoor fire detection, gas leakage, saving electricity, elderly monitoring, kids care, and gesture control [2]. These varieties of applications collectively provide use with a smart indoor system that is helpful in everyday life activities in terms of reducing human workload and revealing fears about home issues or nasty situations. In recent times, and broad proliferation and prominent improvements in sensing methods, Internet of Things (IoT) technologies, and communication techniques have brought us with a diverse set of data about the human especially in indoor environments. This large diversity in data can be collected, cleaned, and analyzed to offer a wide range of indoor services or applications to assist the living of human beings [3].

The development of intelligent IoT applications increases the feasibility of designing a smart system to improve the quality of human lives in an indoor environment. To the end, it is commonly accepted - with minor variations with other work- that the idea of analyzing smart indoor data can be summarized in the five-step workflow as shown in Fig.1: problem formulation, data collection, data preprocessing, data

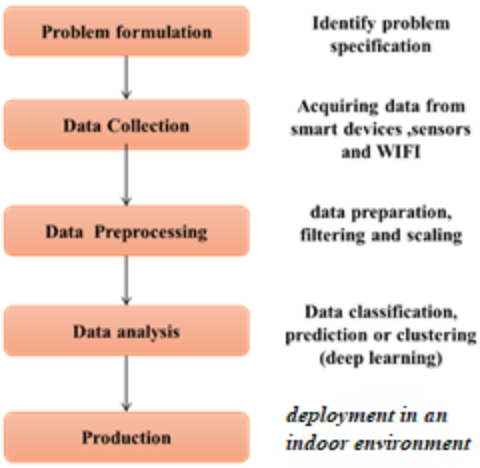

Fig.1.The workflow of human-centric IoT applications in indoor mart environments

analysis, service provision [4]. Firstly, in the problem formulation step, we specify the problem and different aspects to consider. Secondly, the data collection step is responsible for gathering and acquiring various smart indoor data from different devices and IoT sources. Thirdly, in the data preprocessing step, the acquired data is prepared in different ways (e.g., data filtering, segmentation, and scaling) to improve the quality of data for subsequent- steps, as IoT data, often encompasses missing values, noise, uncertainty. Fourthly, the data analysis steps include using mathematical, machine learning (ML), or deep learning (DL) techniques to carry out complex analysis for learning and extracting higher-order patterns and features from previously preprocessed to provide beneficial insights suitable to a diverse set of smart indoor applications. (e.g., elderly monitoring, smoke detection and predicting human fall). Fifthly, the production step is to deploy the output of analysis (DL model) into smart services and applications as a solution for the problem specified in the first layer.

Concerning the steps mentioned above, data analysis is the most significant and vital as it is responsible for processing the data to obtain informative knowledge necessary to make the final decision in any IoT systems [5]. Whereas conventional techniques generally integrate specialized knowledge with ML models (e.g., linear regression, Naïve Bayes, and Decision tree) to perform clustering, classification, or prediction using human-related IoT data. However, the performance of ML models gets declined with high dimensional, heterogeneous, and large-scale data amount, which is often available in indoor scenarios. More, the ML algorithms often require robust feature engineering techniques (automatic or handcrafted) and the performance of algorithms heavily hangs on the efficiency of these techniques.

Recently, deep learning (DL) has been demonstrated as an efficient solution for the limitations of ML approaches and showed excellent performance in different research areas such as natural language processing (NLP), computer vision (CV), speech recognition, and time-series analysis [6],[7]. The primary advantages of DL that lead to this success include 1) DL incorporates deeper neural network architectures capable of extracting sophisticated hidden patterns from the enormous real-time raw data in various IoT devices compared to machine learning shallow models. 2) DL data processing capability is controlled mainly by the depth and the type of learning models. 3) DL can learn effective features from complex raw data without the participation of the inefficient handcrafted feature specification. Motivated by this, the research community decided to take advantage of DL models and the associated powerful learning capabilities to develop more efficient and effective approaches for different human-centered Indoor IoT applications, i.e., indoor localization [8], fall detection [9], Activity monitoring [10], energy control [11], and robotic control [14]. Therefore, this study emphasizes studying the contribution of a different kind of DL to improving the efficiency of IoT applications, focusing on indoor located ones.

The indoor IoT applications can be categorized according to device dependency into two main groups: device-based (device-dependent) and device-free (device-independent) applications. DL model input data obtained from a physical device directly connected to the human beings in device-dependent systems. The device-dependent indoor applications can be divided into two classes based on the sensing modality, particularly vision-based and sensor-based. **Vision-based** approaches are extensively exploited DL capabilities to analyze images and videos captured by a camera in smart indoors for varieties of indoor applications. Yet, these approaches still suffer from some obstacles regarding illumination, power consumption, and computational complexity. By contrast, **Sensor-based** approaches are more powerful in an irregular environment and can facilitate developing a portable system. Meanwhile, the data can be aggregated from a wearable sensor (e.g., data glove), smartphone sensor (accelerometer, gyroscope, and magnetometer), or temperature sensors [12]. However, the requirement of firm attachment and accurate placement make these approaches more incompatible with the nature of indoor life as the devices/sensors might be damaged, lost, misplaced or necessitate regular battery recharges/changes. On the other hand, **device-independent** applications overcome privacy and intrusive issues as they primarily depend on sensing components situated in the indoor environments, which can act as the fundamental data source for developing human/robot-centered applications without necessitating direct contact or attachment to a particular sensor or device. Unlike previous surveys that only focus on one of the before mentioned categories of IoT applications, this study considers surveying the DL studies for both device-dependent and device-independent scenarios [5], [13].

A. Study Contribution

To sum up, this study contributes by providing a comprehensive view of the role of DL methods in indoor IoT applications according to the following points:

- This study presents a thorough overview and categorization of the latest improvements in deep learning methods for human-centered IoT applications in Indoor smart environments. Firstly, we categorize the current studies based on the device dependency (sensing technique). Secondly, the DL models are essentially categorized and contrasted according to the underlying learning strategy. Finally, we provide a taxonomy for categorizing the recent literature from an application domain perspective.
- This study also a tabular review of the publicly available datasets/benchmarks, including vision-based data, sensor-based, radiofrequency and others. The main target is to inform the researchers about the available data that could be readily exploited in evaluating and experimenting with their newly developed DL approaches for human-centered indoor IoT applications.
- We review and analyze DL approaches for different human-centered indoor applications by providing a tabular one-to-one comparison. The methods were compared in terms of the proposed DL model, accuracy, application, system configuration, and data employed.
- This study deliberates the contemporary shortcomings, challenges, and chances for future research of DL techniques for improving the efficiency and effectiveness of human-centered IoT applications aiming to improve the quality of life of humans in an indoor environment.

B. Study Organization

We can summarize the structure of our paper as follows. In Section II, we discuss some related work; In Section III, we provide a background overview for DL models. In Section IV, we compare recent DL models for smart indoor applications in three different domains. In Section IIV, we point out some observations and findings. In Section VI, we discuss challenging issues and future directions. Finally, in Section VII, we provide a conclusion for our study.

II. RELATED SURVEYS

This section discusses the recent survey of indoor IoT applications, including the device-dependent and the device-independent application.

A. Surveys on Device-Dependent Approaches

The multiplicity of device-dependent approaches has been developed for different kinds of IoT applications, especially those based on camera data or sensory data. In this regard, Li et al. [14] presented a comprehensive overview of the computational techniques of multi-individual activity recognition and the relevant applications in a smart IoT environment. They primarily discuss the analysis and fusion of sensory data (vision or sensors) and give some insights into the challenges and opportunities for collective activity recognition. Abuhamad et al. [15] Reviewed around 140 studies of constant human authentication techniques by classifying them into six interactive and physical biometrics classes including, gesture, gait, voice, motion, keystroke dynamics, and multimodal. They also compare the relevant studies based on the sensors, modality, algorithm, and user data. The authors also discuss the intuitions and challenges of the current biometric methods that can be addressed in future work. Dang et al. [12] surveyed and analyzed studies for human AR methods and indicates the corresponding merits and demerits, where the AR methods are divided into two categories i.e., sensor-based and vision-based methods which are compared based on the data aggregation, preprocessing methods, feature extraction, and the training procedures. The authors also treat the human activities at different levels of human-object (HO) interactions, human-human (H-H) interactions, grouping activities, gestures, and actions. Guo et al. [16] argued the fusion-based localization systems and methods from different sources of various networking frameworks, including Homogeneous systems, Heterogeneous systems, and Hybrid systems. However, the device-dependent approaches are obtrusive and not user-friendly as it necessitates the human and device to be attached or located in the same place. They are greatly affected by environmental obstacles.

B. Surveys on Device-Independent Approaches

In an attempt to address the beforementioned limitations of device-dependent approaches, device-independent sensing technologies emerged to enable the design of indoor IoT applications without reliance on attached devices or monitoring cameras. Hussain et al. [13] reviewed the device-free approaches for recognizing different categories of indoor human activities, including action-based activities, motion-based activities, and interaction-based activities. They also provided a taxonomy for this activity recognition task into ten different subcategories. Zhu et al. [17] surveyed the ML and intelligent methods for indoor localization using different kinds of fingerprints and introduced a new framework for intelligent localization. They also discuss the main issues of designing intelligent localization in the real world and accordingly discuss the possible improvements and future solutions. Alam [8] Presented an extensive overview of the non-RF-based approaches for device-independent indoor positioning in the IoT environment. The authors consider light-based studies, infrared-based studies, physical excitation-based studies, and electric field sensing-based studies, then discuss the main limitations of each kind of those studies and the promising research directions. Deep et al. [18] overviewed the research studies of anomalous behaviors for elderly caring within indoor IoT environments while emphasizing dense sensing-based methods due to their robustness to the situational variations, non-intrusive nature and sociability. The authors also provided an overview of the main issues and associations between human activity and anomalous behavior. Nirmal et al. [19] presented a complete review and taxonomy of DL studies for RF-based human sensing by comparing different types of algorithms and offering a more detailed view of the DL for human-centric RF-based sensing. They also reviewed 20 released benchmarks of labeled radio signals of human activities. He et al. [2] investigated the recent innovations in WiFi vision tasks, including the sensing, recognition, and detection from channel state information (CSI) of the commodity WiFi devices. They emphasize using these tasks in 9 essential applications of IoT environments, involving WiFi

imaging, vital sign observing, human identification, indoor localization, gesture analysis, gait analysis, daily AR, fall detection, and human detection. Liu et al. [20] surveyed the wireless sensing approaches in the context of their fundamental preliminaries, methods, and system constructions. Then, discuss how the wireless signals can be used to ease the design of different IoT applications containing indoor localization anomaly detection, room tenancy observing, daily AR, gesture recognition, vital signs observing, and human identification. They also outlined the future opportunities of exploiting wireless signals for human-centered applications. Thariq Ahmed [21] argued the gesture recognition approaches in the context of device-independent sensing based on CSI measurements, which can be categorized into model-based and learning-based approaches. They also discuss data preparations, employed feature engineering and classification models, and the environmental considerations that impact the performance. Furthermore, Zhang [5] surveyed ML- and DL-aided wireless sensing for detecting humans from Red Green Blue (RGB)/Depth imaged and radar data by merging and fusing information from the heterogeneous kinds of sensors as a way to enhance the total performance of realistic human detection approaches.

III. OVERVIEW OF DEEP LEARNING METHODS

DL belongs to a category of artificial intelligence that hires a deeper neural network with many layers of interrelated neurons to extricate pertinent and valuable representations from large-scale and high-dimensional data. Basically, the neurons in each layer get activated to generate a production signal based on a group of weighted values received from the previous layers. Effective learning of such models could be realized by the iterative update of neurons' weights in accordance with the extra training data passed as a network's input. In the past, such DL models were not believed desirable owing to their substantial computation and time requirement. Recently, the great improvements in computational design bring us to more powerful computational agents i.e., graphical processing units (GPUs), and Tensor Processing Unit (TPU) which make the DL development less costly. Accordingly, the research attention moved toward investigating the potential of DL models and their applications in a wide variety of domains.

The massive research in the latest times has generated an abundance of DL models with a wide variety of traits and benefits. These have been generally adopted for various application domains; meanwhile, only some are mainly concentrated pursuing IoT environments. This section presents a concise introduction to the commonly studied DL models, which have been effectively employed for indoor IoT applications. The studies under the umbrella of DL can be categorized according to the learning strategy into three primary categories, namely supervised, unsupervised, semi-supervised, and reinforcement learning models. This section mainly investigates the usages and attributes of different categories of DL models as discussed in subsequent subsections. For additional detailed information concerning the construction and implementation of these models, the readers should refer to original studies of these models. The relevant indoor IoT applications are completely discussed in Section III.

A. Deep Supervised Learning Models

The supervised models are referred to as DL models that are trained to learn the inherent representations from labeled datasets, where the main target is to minimize the difference between the model's estimated output and the actual data labels, which is computed using a predefined loss function.

1) Multilayer Perceptron (MLP)

The MLP is considered the fundamental supervised DL model is comprising of an input layer, an output layer, and single or numerous completely linked hidden layers. Each layer comprises single or multiple perceptrons. The input layer is responsible for receiving the input data, where the number of input perceptrons is mounted to cope with the size of the input vector of the underlying problem. The hidden layer's neurons receive received input weights, then get activated by a non-linear function to generate output values, which are followingly passed to the subsequent layer (forward propagation). The training

procedure happens incrementally by renovating the learned parameters after each training batch is handled, according to the difference (i.e., loss) between the estimated and the actual output (backward propagation). The activation function of the output layer is usually determined by the based type of problem (prediction, classification, etc.) [5].

2) Convolutional Neural Network (CNN)

The CNN one of the most widely used DL models that achieve powerful results in many research fields, especially computer vision and pattern recognition, because of its ability to capture and learn the sophisticated pattern in multi-dimensional data. Particularly, CNN hires a stack of convolutional filters that convolve over the input vectors to extract spatial representations. Also, it employs different form pooling layers to lessen the dimensionality of generated feature maps during training. Remarkably, even though CNN initially developed to process the visual data (i.e., image, videos, etc.), it is also observed to be efficient in capturing the spatial interactions in sequential data, i.e., textual data, voice data, time series [22].

3) Recurrent Neural Network (RNN)

The RNN is designed based on FFNN architecture for modeling sequential information using a kind of feedback structure called recurrent unit, which enables memorizing the temporal relationships within the input. Unlike other FFNN, output at any time step depends only on the present input with no dependence on previous input/output. Model memorizes the previous output state and uses it as an input in conjunction with the current input, which makes RNN extensively used for time series and sequential data like speech recognition, sensor data natural language processing. Nevertheless, the RNNs exhibit two critical drawbacks. First, the gradient vanishing problem happening because of trivial gradient updates that end the learning process. Second, gradient explosion happens due to growing weights' gradients through the backward propagation, which causes a big gradient update [23]. These shortcomings limit the RNNs capabilities from modeling long-term dependencies. In order to tackle these limitations, long and short time memory (LSTM) has been designed to effectively regulate the addition or removal of memory state using a simple gating mechanism (input, output, and forget gate) that optionally remembers or forgets the information. The main issue with LSTM is the huge number of parameters [5]. This motivates the design of another variant of RNN called gated recurrent unit (GRU), which is a simplification of the LSTM network by combining the input and output gate in a single gate, the so-called update gate. This makes GRU a lightweight and more efficient model for modeling long-term dependencies while maintaining the simpler architecture. The idea that learning in the current time step does not depend only on the historical information but also on the future information motivated the design of bidirectional RNNs (Bi-RNNs) that process the input sequences both forward and backward directions. This encouraged the usage of bidirectional LSTM (Bi-LSTM) and bidirectional GRU (Bi-GRU), which have been showing great success for a wide variety of applications [19].

B. Deep Unsupervised Learning models

The unsupervised training of DL models is referred to the situations where the DL models are trained to learn the inherent representations from an unlabeled set of data. This is usually beneficial under circumstances where data labels are unattainable, or the data labeling/annotating is very laborious and time-consuming. In the following, we overview the most common unsupervised models.

1) Auto-Encoder (AE)

An unsupervised DL model is basically used to reduce the dimension of data by performing two consecutive tasks, which are encoding and decoding. The encoding operation aims to learn complex implicit features from raw input data while compressing it into the lowest dimensional representation called latent representation. In the decoding phase, the AE seeks to reconstruct the original data from the compact latent representation [24]. Since the input operates like an output, the AE is able to act in a self-supervising manner without demanding specific data labeling. The AE has been widely used as an efficient dimensionality

reduction tool and has shown great feature extraction capability in hybrid DL models. This great success inspired the development of different variants of the AEs network. For example, stacked AE (SAE) extends the AE by including a sequence of multiple cascaded hidden layers between input and output layer to extract hierarchical feature increasingly from data providing fine-grained data representation as an output. The Sparse AE seeks to realize information bottleneck using sparsity conditions which alleviate the need for reducing the number of nodes in the hidden layers. Instead, it penalizes per layer activations during the encoding and decoding process [25]. Besides, the De-noising AE (DAE) is specifically developed to produce a clear and finetuned output based on extracted features from distorted or partially corrupted data (i.e., noisy images). The Variational AE (VAE) is an improved variant AE devised to learn and describe an input in latent representation probabilistically. In other words, it employs the encoder to define a probability distribution for each latent variable rather than generating a single value for portraying each latent variable [26].

2) Restricted Boltzmann machine (RBM):

An unsupervised DL was designed as an extension of the Boltzmann machine (BM), where BM consists of circular unidirectional connected nodes trained to determine the nodes to be activated. To overcome the overhead incurred by the massive connection among the node makes nodes of BM, the RBM is introduced to divide the nodes into two layers (i.e., input layer and one hidden layer) and eliminate the connection among nodes of the same layer (intra-layer connection). The RBM has been gained a wide adoption for feature extraction, parameter initialization, and collaborative filtering.

3) Deep belief network (DBN)

An unsupervised DL model designed by stacking multiple RBM networks, where each hidden layer of an RBM is considered an input to the next RBM with symmetric connections between them. Except for the foremost and last layers, every layer of DBN has two main roles. First, it acts as a hidden layer with regard to the previous layers. Second, it acts as an input layer for the subsequent nodes [27]. The DBN is trained in a greedy layer-wise way with optimized weights to conceptualize the learned features drawn from the original data. The main goal is to realize faster-unsupervised training by regulating the weights based on contrastive convergence of constituting RBM to create a stable approximation of the possible scores. The DBNs have been achieving great success in recognizing, generating, and clustering video streams, images, and action-catch data [28].

4) Generative Adversarial Network (GAN)

The GANs is another example unsupervised DL model intended to acquire the distribution of training data. Unlike VAE, the GAN concurrently trains a couple of networks (i.e., generator and discriminator) to act like a two-player game (hence called adversarial) to create observations without acquiring the circulation parameters promptly. The generator seeks to generate fake observations similar to genuine observations and send them to the discriminator. The discriminator is simultaneously trained to recognize the generated observations as fake and compute a probabilistic value, representing data authenticity to penalize the generator for improbable data. The generators and discriminators are optimized throughout the training until the discriminator becomes unable to differentiate the real observations from the generated ones. GANs have definitely transformed the DL research with several alternatives of GAN architectures [29].

C. Deep semi-supervised learning models

The spectrum of semi-supervised DL models includes the models that seek to exploit both unlabeled and labeled data instances during the training process. The labeled data offer partial supervision, and the unlabeled data delivers valuable indications concerning the data distribution, which can empower the model capabilities to learn more improved data representations. For instance, an efficient DL model must produce steady and soft estimations under arbitrary perturbations (e.g., scaling, rotations, translations, flipping, or accidental trepidations) GAN [30] alongside the input space or prevent retaining the judgment factors on high-intensity regions of input space distribution. The current semi-supervised models can be reviewed from two distinct viewpoints, namely generative models and Teacher-Student models.

1) Generative models

the semi-supervised AEs, RBMs, DBF, and GANs could be obtained from the equivalent unsupervised DL models. For example, the semi-supervised GAN gets trained a K+1 classifier with K-provided data labels as well as a label of fake samples. Then discovers the distribution of unlabeled observations by handling them as a subset of the first K genuine classes, and the competitive capabilities are released via feature matching technique [31]. Besides, the semi-supervised AE train a classifier for predicting from the latent representation encoded from the labeled and the unlabeled subset of the training data. In contrast, the classifier uses the labeled observations.

2) Teacher-Student models

The Teacher-Student models are regarded as the type of semi-supervised models that have been realizing great success in recent years, in which one or more teacher networks are trained to estimate the labels with for unlabeled observations. The estimated labels are employed to regulate the parameters of a student network during training. The constancy between the teacher and the student should be boosted to enhance the capabilities of the student network in classifying the unlabeled observations [32]. The Teacher-Student models have transformed the research in semi-supervised training with a variety of models, including Noisy Teachers, Temporal Ensembling [33], Mean Teacher[34], Mixture-Match Teachers [35], Fix-Match Teachers [36], and Adversarial Teachers [37].

IV. DATA COLLECTION AND BENCHMARKS

The DL research community is witnessing an increasing demand for open-source datasets to encourage reproducibility and accelerate research production. The crucial target is to deliver a wide variety of benchmarks to easily experiment and contrast the performance of DL models from different and autonomous studies. Unfortunately, the aggregation and annotation of the human-centered dataset is an exhaustive task. The data available to the public is not viable every time, especially indoor data, because of privacy limitations and maintaining datasets might be expensive. Though most researchers are presently accumulating their datasets in the lab to experiment with the proposed DL algorithms, gaining access to community/released datasets is potentially ideal for speeding up the deep learning research in the future. Thus, this section extensively overviews the publicly available benchmarks for different indoor IoT applications and highlights the current data deficiency for those applications that might suggest exploration in future research.

A. Activity Recognition datasets

1) Vision datasets

The vast improvements and the wide applicability of computer vision models attracted the research attention toward the design of innovative AR approaches based on visual datasets. In view of this, the vision-based AR approaches could be categorized according to the type of data to involves RGB data [38], [39], [40], and RGB-D data [41], [42]. Commonly, the RGB-D-based AR approaches result in better performance than RGB-based approaches due to the extra information provide by multi-modality and depth data. Nevertheless, the RGB-D exhibits high computational and design complexity and excessive expenses, making the RGB data dominating the design DL-based AR approached. Table I presents the characteristics of current vision-based AR datasets in terms of the type of AR, number of subjects, number of classes, number of samples, the presence of depth information, and the number of clips per class. A short description of the nature of recorded activities is outlined along with the source of data.

RGB data: An RGB image comprises red, green, and blue channels in the observable continuum that might be filmed utilizing standard cameras. It is broadly accessible, inexpensive, and offers abundant texture data

about the human. Nevertheless, the cameras often have a constrained scope, prone to calibration, and are severely affected by environmental settings, i.e., such as walls, illumination, and lighting.

RGB-D data: The emergence of the depth sensors and scope visioning methods brings additional valuable information that can improve the capabilities of learning algorithms in recognizing human behaviors and actions. Besides, the skeleton data could be captured from the depth information to offer a dense interpretation of the human skeleton. The low dimension of skeleton data accelerates the learning of DL models. Hence, abusing the joint information obtained from measurements of depth sensors becomes a desirable research path as it could be applied in various IoT applications. The RGB-D data has several merits, including the robustness against variations in illumination, lighting, colors, and textures, and pitch-black situations. Nevertheless, RGB-D data exhibit a limited resolution, instituting noise to the images owing to quiet compassion, and could be effortlessly disturbed by light-grasping and translucent materials.

Table I. An overview of vision-based indoor activity recognition datasets.

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

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

2) Sensory data

With the swift progress of wireless sensing, a large number of sensory measurements can be captured from a variety of inexpensive and widely available smart sensors to facilitate the development of DL models for recognizing human activities in smart indoor environments (i.e., smart home and smart healthcare). The current sensory data can be categorized according to the sensor modality into four categories of sensors,

namely Ambient sensors (AS), object sensors (OS), Wearable sensors (WS), and hybrid sensors. Table II presents a detailed description of four sensor modalities, common sensors, the corresponding data, merits and demerits.

Table II. An overview of characteristics, merits, and demerits of different categories of indoor sensors.

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 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 for only 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 throughout 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 torquers.
Hybrid sensors	This refers to the studies that employ a different combination of the beforementioned sensors modality to improve the efficiency of indoor IoT application by empowering the representational capabilities of the DL model.			

Moreover, Table III reviews the publicly released benchmarks that are typically employed to train the DL models for the sensor-dependent approach. It is notable that the majority of the current benchmarks are aggregated using WSs, including magnetometer, gyroscope, and accelerometer. Table III also presents the number of observations, activities, subjects, and attributes in each dataset. Among the tabulated data, the UCI datasets (OPPORTUNITY[63], HAR [64] , M-HEALTH [65] , etc.) and WISDM datasets are the commonly used as a standard datasets for evaluating the sensor-dependent deep learning approaches. It is also observable that most of the present sensory datasets are aggregated by capturing the activity of single individual, meanwhile few of them consider multiple or group activity. Regrettably, the obligation of carrying a device or some sort of sensors is burdensome and potentially impractical for humans, so, making it difficult to develop ubiquitous IoT applications especially in indoor environments.

Table III. An overview of sensor-based indoor activity recognition datasets

ID	Dataset	Type	#Subj	#Act	#Attr	#Obs	Devices	Sensors	Sampling rate
SA1	WISDM 1 [66]	Single	29	6	6	1,098,207	Sw	A	20 Hz
SA2	WISDM 2 [67]	Single	36	6	6	2,980,765	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	120	WS	A, G, M	50 HZ
SA8	HHAR [70]	Single	9	5	16	43,930,257	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	20,750	Sp	A, G	100Hz
SA12	PAMAP2 [74]	Single	9	18	52	2,844,868	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	598,396	Sp	A, M	50 Hz
SA20	ARAS [82]	Multi	4	27	21	5,184,000	WS	A, M	10Hz
SA21	CASAS [83]	Single /Multi	2	15	NA	NA	AS	AS	NA

A=accelerometer, G=gyroscope, M=magnetometer, OS=object sensor, WS=wearable sensor, AS=ambient sensor, Sp=smartphone, Sw=smartwatch, IMU=Inertial Measurement Unit, NA=" not exist"

3) Radio Frequency data

The privacy-preserving, contactless and non-LOS characteristics motivated the researchers to exploit RF signals to develop an intelligent IoT application. In RF-based applications, the transmitter broadcasts the RF signals, preceding the arrival to the receiver, they get regulated by the human beings and their indoor behavior. The regulated data can be collected, cleaned, and analyzed using DL models for different purposes based on the kind of application. In this regard, the RF data can be divided into three main categories based on the communication technology employed. First, Radio frequency identification (RFID) is a favorable communication technology that alleviates the need for device attachment (sensor) and has many benefits. In Essence, the RFID utilizes electromagnetic fields to instinctively recognize and track the tags affixed to different entities, which encompasses electronically collected data. The RFID tags can be active or passive tags, where the former ones depend on a small power supply to constantly transmit the detectable RF waves from hundreds of far away distances from the RFID reader. Conversely, the latter tags accumulate power from a neighboring RFID reader probing RF signals to deliver its collected information. Hence, the passive tags of the RFID system are much inexpensive and softer. Second, radar-based, which is considered an active sensing technology in which the RF waves are broadcasted, then modulated by target and get received in a modulated form. Radar technology has been widely used for outdoor applications such as traffic control and remote sensing. Recently, there is a growing interest in exploiting radar technology for indoor applications (i.e., indoor positioning, navigation, activity recognition, etc.) due to its contactless nature, simple construction, easy deployment, comparatively cheap, etc., inattentiveness to weather and lighting state, and penetration ability. To this end, two kinds of radars are currently available, namely Continuous-Wave (CW) Radar and ultra-wideband (UWB) radar. The CW radar broadcast a notorious constant-frequency CW ratio signal and accepts the moderated signals by targets (human) on the pathway of the signal. The CW radars can operate in either moderated or unmoderated manner, and it includes frequency-modulated CW (FMCW) radar, and interferometry radar, and Doppler radars. Third, the WiFi networks are the most widely used RF technology, mainly in indoor environments because of their cheapness, rapid transmission rate, and handy installation. This result is a new form of data that can be exploited for designing human-centric applications, including Received Signal Strength (RSS) and channel state information (CSI), where the RSS denoted the mean of amplitude information throughout the entire channel bandwidth. Meanwhile, the CSI encapsulates the frequency reaction of the wireless channel, denoting the

way the phases of various frequencies get changed and attenuated throughout the broadcast from the transmitter to the receiver. Table IV presents the characteristics, measurements, merits, and demerits of different RF communication technology. A wide variety of Human-centric applications can be developed based on the beforementioned technologies by extracting different sets of RF measurements, including the Angle of Arrival (AOA), Time of Arrival (TOA), Time Difference of Arrival (TDOA), and other similar. The downside of such techniques is the low accuracy and high energy consumption.

Table IV. An overview of characteristics, merits, and demerits of different RF communication technologies.

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 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 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		

The sensitivity of RF data to the device configurations and experimental conditions has been exhibiting many troubles in evaluating and comparing various DL studies. Mercifully, some research studies have open-sourced their aggregated RF datasets, offering an opportunity for other researchers to reinvestigate, analyze these data, and reuse them to evaluate and compare different DL algorithms fairly. Table V display and survey the current datasets in terms of AR level, number of participants, number of classes, number of observations, a sensing device, RF signals, and activity description. The reviewed benchmarks lead us to some critical insights, which can briefly be discussed as follow. 1) the daily activities, gestures, and gait analysis dominate the current datasets, while other applications such as respiratory monitoring, human counting, fall detection still have a limited number of benchmarks. 2) the number of subjects varies from one subject to 95 subjects, which provides an indication of the limited diversity in the dataset. 3) The number of environments considered during the aggregation of datasets is ranging from one to seven, necessitating a greater number of distinct indoor environments to increase the variability in data and the generalizability of DL models. 4) The CSI is the most common RF signal used for AR and always captures with Intel 5300 Network Interface Card (NIC), which indicates its precious value for modeling different human activities. 5) the vast majority of the RF datasets for AR are accumulated for single user activity and

only one of them [84] considers multi-individual activities. Besides, the data presented in [84] only consider the interactions between pairs of subjects, making the group activity RF data still unavailable.

Table V. An overview of RF-based indoor activity recognition datasets

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	6	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	15,635	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"

B. Indoor positioning and tracking datasets

In the same way, the publicly released datasets for indoor positioning fall in one of three categories, i.e., vision data, sensory data, or RF data. Table VI reviews the different characteristic public indoor poisoning benchmarks in terms of dataset name, type of data, number of subjects, number of attributes, number of observations, devices included number communicators, frequency bands, and the data measurements. It is observable that the RF data (CSI, and RSS) dominate the present benchmarks because of its high efficiency for modeling positional information of the humans compared to the sensor or vision-based data. Unlike activity recognition benchmarks, there are a limited number of indoor positioning benchmarks and most of them are legacy data.

Table VI. An overview of vision-based indoor positioning and tracking datasets.

ID	Dataset	Type	#Subj	#Attr	#Obs	Devices	#devices	Frequency bands	Data
IL1	Jekabsons [98]	RF	1	20	82P/68P	RBT-1002, RBT-4102	14 APs	2.4 GHz, 5 GHz	RSS
IL2	ARIL [99]	RF	1	2	1440	EttusN210	2 antennas	2.4 GHz or 5 GHz	CSI
IL3	IPIN 2016 [100]	S RF Non-RF	>1	>=1	NA	Smartphones Samsung, Sony, Hawaii	>1	50 Hz <10 Hz 0.25 to 0.17 Hz	RSS, A, G, M, sound, light data
IL4	UJIIndoorLoc[101]	RF	>20	529	21,049	Android Smartphone	520 APs	NA	RSS

IL5	UJIIndoorLoc-Mag [102]	S	NA	13	40,195	Google's Nexus 4 LG G3	2 SPs	10Hz	A, O, M data
IL6	IPIN 2016 Tutorial [103]	RF	NA	177	1629	NA	168 APs	NA	RSS, GPS data
IL7	ALCALA 2017 [103]	RF	NA	154	1075	NA	152 APs	NA	RSS, GPS data
IL8	crowdsourced[104]	RF	8	992	4648	21 android devices	991 *APs*	2.4-GHz and 5-GHz	RSS, GPS data
IL9	Widar 1.0 [105]	RF	5	30×3	NA	3 mini-desktops	1AP	5.825 GHz	CSI
IL10	Widar 2.0 [106]	RF	6	30×3	NA	2 labptop	1AP	5.825 GHz	CSI

S="sensor", SP="smartphone", NA=" not exist"

V. INDOOR IOT APPLICATIONS

A smart indoor enables the interconnection of pervasive IoT devices embedded in many indoor appliances such as smartphones, smart television, smart fridges. Recent advances in DL motivate the researcher to use DL to address many smart indoor issues that help to enhance life quality with different applications of smart indoor environments. This section reviews various DL approaches for different categories of human-centered IoT applications in smart indoor environments.

A. Positioning and Tracking

Identifying human locations has been shown great importance for human-centered applications. Location-based services (LBS) describe the services designed to attain the physical location of individuals using different localization approaches. Despite the success of outdoor positioning/ navigation technologies and the broad adoption in the daily lives of human beings, they fail to keep performing well in the indoor environment because of its complexity and the associated environmental conditions. Hence, the real-time and efficient identification of indoor positions becomes a vital and challenging research area for future smart buildings. Generally, the indoor localization DL approaches could be categorized into three groups: vision-based, sensor-based, and RF-based.

1) Vision-based Indoor positioning

The CV techniques are known to be robust and reliable for a wide variety of applications. In this context, Zhao et al. [107] investigated the viability of improving the localization performance by combining the camera data with smartphone data and WiFi data to develop a multimodal framework that can help to reconstruct the inner vision of building for subsequent positioning or navigation. Ha et al. [108] presented a novel visual indoor localization framework that integrates CNN with building information modeling (BIM) to create a benchmark of condensed BIM images and then explored the data to find the utmost analogous to indoor photographs, in that way approximating the indoor location and direction of the photograph. Another study [109] employed CNN for indoor positioning of unmanned Aerial Vehicles (UAV) based on transfer learning design, where the genetic algorithm is employed to optimize the model's hyperparameters. However, visual data is not the most suitable option for determining the local coordinates and distances. Also, the vision data are known to be privacy obtrusive and obey the LOS restrictions. Thus, the vision-based approaches become the non-preferred choice for indoor positioning, tacking, and navigation.

2) Sensor-based Indoor Positioning

To address the limitation of vision-based indoor positioning approaches, the researchers believed to exploit the sensory data generated from different sensors embedded in smartphones, smartwatches, etc. For example, the authors of [110] presented a DL framework for indoor localization from geomagnetic data captured by magnetometers, then encoded the geomagnetic data into recurrence plot representations, which are followingly fed into CNN for automated feature extraction and later for classification. In [111],

the authors designed a DL system that uses LSTM to learn to perform indoor positioning based on bimodal sensory information, including data sensed by light sensors and magnetometers. The experimental evaluations on private datasets validated the practicality and feasibility of using bimodal data in improving the localization performance. For exploring more sensor modalities, the authors of [112] introduced a multi-sensor DL framework that considers learning from multimodal data from light sensors, magnetometers, barometric sensors, pressure sensors, and Global Navigation Satellite System (GNSS). The framework employed dense convolutions for efficient feature extraction, used LSTM to model temporal dependency, and MLP to compute the classification decision. Distinctively, the Pedestrian Dead Reckoning (PDR) is recently considered one of the typical approaches for achieving indoor localization as a result wide availability of smart devices. In this regard, the SAE network is presented in [25] to estimate the step length in the PDR system using smartphone data (Accelerations and gyroscope data).

In a nutshell, the selection of DL model and sensors to be used for localization heavily depends on various considerations, including the efficiency, localization environment, computational resources availability, latency issues, etc. For instance, efficient localization can be attained by using acceleration, gyroscope sequences, magnetometer sequences. However, carrying or wearing the sensors all time might be extremely problematic for humans. In multi-floor environments, the movement of humans between different floors is indispensable; hence, being aware of the floor level might improve the positioning or tracking performance. Besides, the importance of spatial information or temporal dependency within the data necessitates deciding the appropriate DL layers, i.e., convolutional, recurrent, attention layer.

3) RF-based indoor positioning and tracking

The RF-based indoor localization can be divided into two subcategories, namely fingerprinting and triangulation [12]. The fingerprinting approaches entail the online and offline phase CSI information. Where the system calculates the CSI reports throughout the offline phase at the targeted positions to construct the fingerprint benchmark, meanwhile the gauged CSI measurements are contrasted with the fingerprinting data to decide or track the position of the goal throughout the online phase. On the other hand, the triangulation/geometric approaches decide and track the position of humans depending on the triangles' geometric attributes. Based on Table VII, it is notable that DL models based on RF signals have gain wide adoption compared to the visual or sensory data.

Fingerprinting: The fingerprinting approaches belong to one of two distinct groups, the deterministic and probabilistic approaches [95]. The former approaches employ various similarity measures (cosine similarity, Manhattan, Euclidean distance, etc.) to contrast the estimated data with the original fingerprints to decide the position of humans [73]. For instance, Wang et al. [113] extracted phase information of CSI data to approximate the AoA data, construct AoA images (with size 60×60) and pass them to a deep CNN to be trained for indoor localization during the offline phase. The Latter approaches usually depend on the statistical estimates (i.e., mean absolute error (MAP), mean square error (MSE), root MSE (RMSE)) to compare the model outputs with the original fingerprint, and thereby enable modeling the uncertainty through a different form of the RF data [95]. This implies predefining and storing the information about the distribution of signals across reference points. In [114], the authors integrated the FFNN and decision tree algorithm in a single framework to model the location information from the RSS data, where fuzzy learning is employed to model uncertainty during the training.

The probabilistic approaches are shown to have a variety of features, evaluation criteria, and localization accuracy. However, the most remarkable about them is that the high localization efficiency requires a complete inspection of the environment to construct complete fingerprints and necessitate upgrading the

fingerprints based on environment alterations. These conditions limit the deployment in the real-world indoor environment. Thus, crowdsourcing persons' measurements to construct fingerprint testbed is likely to be a viable solution. DRL is suggested to be employed as a second solution where reward and punishment can be determined according to the environmental changes. In addition, it could be seen that the deterministic fingerprinting localization (Table VII) attain wide adoption in recent studies and achieve promising performance. However, the process of creating a fingerprint database is composite and time-consuming.

Triangulation: The triangulation can be realized in two viable methods, including lateration and angulation. The Lateration method seeks to approximate the target's position by computing the corresponding distance, such as the TOF, with respect to reference points. Xue et al. [115] employed LSTM architecture to learn from erroneous TDOA measurements of UWB signals without degrading the localization performance. He et al. [116] employed a DNN model to learn indoor locations from both RSS fingerprints as well as TDoA measurements. Angulation methods approximate the target's location by calculating the corresponding direction, such as the AoA, concerning the reference points. Wang et al. [117] employed DNN to design a cooperative autonomous generation approach for indoor localization of unmanned Aerial Vehicles (UAV not for humans) using AoA data, where the offline/online application regulations are employed to estimate the optimal heading angles for UAVs. Several studies have been investigated the effect of modeling AoA measurements in improving the performance of indoor localization with around 20% [118], [119]. However, the estimation of AoA mainly relies on the spatial variation in the sensing signals. Therefore, the key reason that restricts the performance of AoA valuation is the number of sensing parities. Therefore, the standardization of new protocols for a huge number of antennas (transmitters or receivers) in the huge MIMO systems will enable better approximation of AoA estimation and improve the dependent indoor localization, navigation, and tracking.

Table VII. An overview of deep Learning studies for indoor Positioning and Tracking

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 multi-view 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	1) An online DL framework for passive human localization by learning the 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

[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 feature 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 icorrect or missed measurements during asynchronous localization called DeepTAL.
[116]	DNN	SU	T	Kalman filter, Distribution Judging, remove 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.

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"

B. Activity Recognition

Activity recognition (AR) has been attracting a growing research interest since it offers an efficient solution for modeling human-computer interactions in smart human-centered systems, thereby bring numerous advantages to surveillance, impaired people care, healthcare systems, etc. With the continuous advancement and affordability of IoT devices and smart technologies, AR becomes a vital task for enhancing human lives in smart indoor environments. Human activities are known to be deliberate, mindful, and personally significant series of acts that can be performed by single or multiple individuals, which might be related or unrelated. Accordingly, the task of AR can be categorized based on the complexity of activities into three main levels, namely individual activity recognition (IAR), group activity recognition (GAR), and hybrid activity recognition (HAR). In the IAR, the main target is to detect and identify the

activities performed by a single subject without considering activities of other subjects in the environment, and this is the lowest degree of complexity of AR. Besides, in the GAR, two or more subjects simultaneously perform some activities, which can be related or unrelated to each other. The related group activities imply that the subjects share the same activity to achieve a joint goal associated with each one of them. For example, when several individuals will pick up a heavy item from the floor to a small table, they owe to cooperate with each other to accomplish this task. On the other hand, the unrelated group activities mean that the subjects perform some actions that are autonomous and irrelevant to other's actions. As an example, when some persons are studying, resting, watching television, etc., and the activity of everyone is independent of the others [126]. Moreover, in the HAR, the main goal is to recognize the individual and group activities in indoor environments, which is a more complex task. For instance, a residential smart home contains three individuals, where a couple of them are washing dishes in the kitchen while the other individual is sleeping. Hence there are a mixture of individual and group activities is being performed in this place. To this end, the research community determined to take the advantages of deep learning for designing efficient and fully automated techniques for recognizing different kinds of human activities using various data modalities as previously discussed.

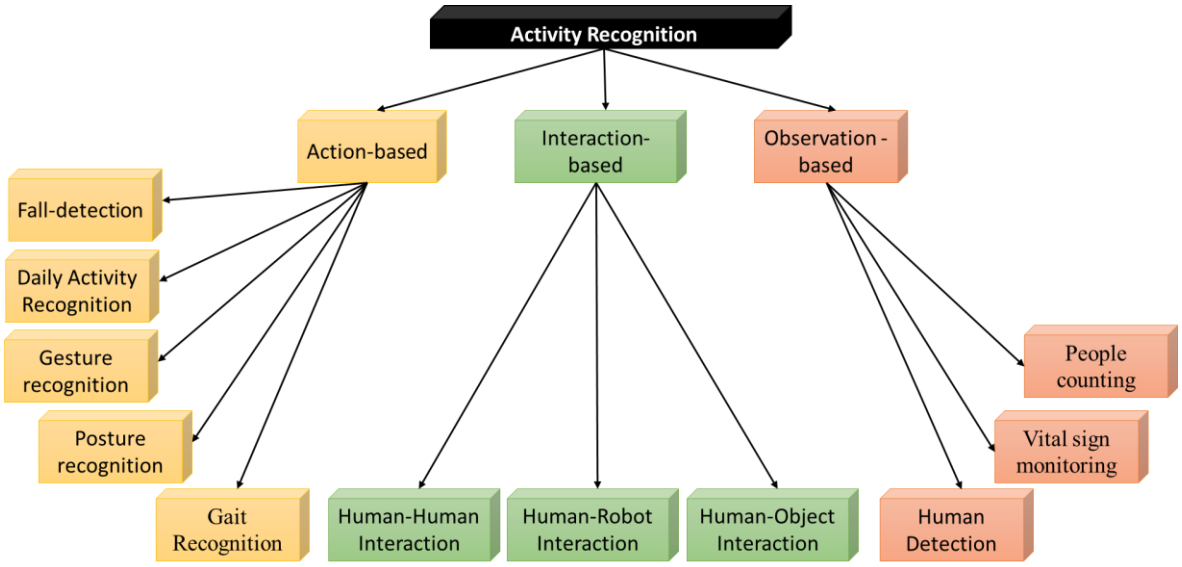

Fig. 2. Taxonomy of classifying the human-centric activity recognition in smart indoor environments.

The AR task primarily seeks to detect and recognize human physical activities either from one or multiple subjects. However, human activities are numerous, spanning multiple categories and classes. For example, some activities require whole-body movements, i.e., running, walking, and lying. Some others required moving single parts of the body, i.e., hand gestures. Other activities are obtained as a result of interaction with other entities in the surrounding environments. To this end, this study presents a taxonomy for classifying the DL-based indoor AR based on the type of the targeted activities. To be more specific, the current DL studies for AR falls in one of three classes, namely action-based AR, interaction-based AR, and observation-based AR. Then, the systematic categorization AR for each class is presented in Fig. 2.

Firstly, Action-based AR considers recognizing the performed by doing some physical actions using either the body or a certain organ. This study systematically divides action-based AR into five main subcategories: fall detection, daily activity recognition, Gesture recognition, posture recognition, and gait recognition. Secondly, interaction-based AR mainly concerns the activities resulting from the interaction between humans and the surrounding entities in the indoor environment. Accordingly, the interaction-based activities can be divided into three main subcategories, including Human-Human Interaction (H-HI),

Human-Robot Interaction (H-RI), and Human-Object Interaction (H-OI). The third class of activities is not associated with any actions or interaction, yet it relates to observing or detecting the presence or absence of humans or observing some aspects of humans. This is greatly beneficial, specifically in security, surveillance, and healthcare application in indoor environments. The observation-based AR include the subcategories of AR, including human detection (Identification), people counting, and vital sign monitoring. The subsequent discussions about the AR consider this taxonomy to survey the DL research presented for these subcategories, emphasizing vision-based, sensor-based, RF-based, and Non-RF-based approaches.

1) Vision-based activity recognition

Vision-based DL approaches mainly depend on the visual sensing tools for monitoring and recording different kinds of human activities in Indoor environments [2]. The most remarkable about this approach is their great dependency on the quality of the captured images or recorded videos. In other words, the resolutions, lighting conditions, illumination alterations, and similar graphical elements are the main factors for determining the quality of visual data. In view of this, the computer vision researchers strive harder to propose a new solution for improving the performance of AR from visual data using reasonable computational overhead to cope with the needs of IoT environments. Table VIII overview the recent DL work on activity recognition in terms of DL model, learning strategy (LS), the AR level, the data preparation (preprocessing and/or feature engineering), dataset identifier, reported accuracy, and contributions.

Table VIII. An overview of deep Learning studies for vision-based activity recognition

Ref	Model	LS	Level	Preparation	Dataset	A (%)	Contributions
[127]	GNN	SU	HAR	- Difference operator - Exponential maps	VA10	96.4	A multi-branch multi-scale GNN to learn to extract spatial-temporal features for AR and motion prediction
[128]	CNN+Bi-LSTM	SU	HAR	- Keyframe extraction - Depth-Human pose model	VA10	84.1	1) A DL framework for AR via spatial-temporal Dynamics (STD) and motion stream. 2) Inception V3 employed for feature extraction in motion stream. 3)The STD employs LSTM and Bi-LSTM to respectively model motion data and temporal patterns of human shape dynamics.
[129]	CNN+AE	SU	HAR	- Keyframe selection - feature descriptor	VA5	96.6	A lightweight CNN-based framework for segmenting multi-view videos into snapshots, then calculate joint information calculation to generate a video summary.
[130]	GCN	SU	HAR	- Resizing - Cropping - Unifying lengths	VA10	95.9	A DL framework employs GCN to learn spatial-temporal features of skeleton information that act as a complementary of RGB feature to empower the learning of action-related information.
[131]	LSTM+AE	SU	HAR	NA	VA9, VA10	92.33, 92.25	A sample fusion model that employs multi-scale transformation architecture for efficient data augmentation for improving the AR performance.
[132]	LSTM	SU	GAR	- Tracklet detection - AlexNet feature extraction	VA22, VA25	98.33, 83.75	A Hierarchical Long Short-Term Concurrent Memory for modeling multi-individual interactions for AR by learning the dynamic correlated features from human interaction clips.
[133]	LSTM	SU	HAR	- Static features - Temporal features	VA25	94.9	A graph LSTM-in-LSTM for simultaneous modeling of the group and individual activities from videos using graph LSTM and person LSTM, respectively.
[134]	LSTM	SU	GAR	- Tracklet detection - person-level features	VA25	93.0	- Employ a Global Context Coherence (GCC) and spatial-temporal Context Coherence (STCC) constraint to seize the pertinent actions and estimate its donations to the group activity. - A Coherence Constrained Graph LSTM presented to model the discriminatory representation of human motions while ignoring the unrelated actions.
[135]	CNN+ LSTM+ GCN+ TCN+ Attention	SU	GAR	- Generate tacklets - Graph construction	VA25	95.8	A teacher-student framework that leverage the previous semantic knowledge, where the teacher learns judicial patterns of various persons via two modules of semantics-saving attention. The learned knowledge forwarded to the student model that obligated to imitate the teacher model.
[136]	GCN	SU	GAR	- Generate tacklets - CNN features	VA25	93.0	A Hierarchical Graph Cross Inference model that combines multilevel information and spatiotemporal relationships between body organs and persons using a cross inference block.

LS="Learning Strategy", SU="Supervised", US="Unsupervised", SS="Semi-supervised", A= "Accuracy"

2) Sensor-based activity recognition

With the recent advancement and wide spreading of sensing technologies, sensor-based AR has become a more attractive research area for DL communities. The sensor data is more lightweight, thereby requiring less computational overhead. Table IX summarizes the recent DL contribution for modeling different kinds of human activities from the sensory data. This includes the DL model, LS, the AR level, the data preparation, dataset identifier, reported accuracy, and contributions.

It is notable that most of the present studies mainly adopt supervised training for their model. A few of them consider learning from unlabeled sensory data, which is large enough and easy to obtain. The authors of [137], [138], and [139] tried to address this issue via semi-supervised training where the limited amount of labeled data involved during the training along with a large amount of unannotated data. Alternatively, the authors of [140] presented an online learning LSTM model for unsupervised training from the unlabeled sample, where the hierarchical k-medoids clustering (Hk-mC) algorithm was employed for automatic labeling of raw signals and building a hierarchical classification. Nevertheless, the potential of unsupervised/semi-supervised DL still not fully explored AR. Besides, the vast majority of the reviewed studies emphasize the AR in the level of IAR. Exceptionally, the authors of [141] tried to address GAR by applying temporal convolution network (TCN) and LSTM network to model multi-user activities from custom 2D Light detection and ranging (LiDAR) data.

Table IX. An overview of deep Learning studies for sensor-based activity recognition

Ref	Model	LS	Level	Preparation	Dataset	A (%)	Contributions
[142]	CNN	SU	IAR	Augmentation	SA2	95.7	A recognition method that uses two-stage CNN to learn from augmented sensor data
[143]	NN	SU	IAR	Segmentation, feature selection	Custom	91.78	New online update phase to estimate true labels for the new data
[144]	CNN	SU	IAR	Segmentation, Pixel encoding, handcrafted features	SA5, SA10, SA13,	97.4, 97.2, 96.1,	1) Fusion model for combining handcrafted and convolutional features. 2) Encode sensor data into pixel values.
[137]	CNN	SS	IAR	a low-pass filter, coordinate change, normalization, DAE	SA14	95.81	1) New Uncertainty-aware multiple-domain CNN framework; 2) A transfer learning employed from synthetic to real data to improve the recognition
[145]	LSTM	SU	IAR	Normalization	Custom, SA6, SA15	85.6 93.5 80.3	A cost-effective multi-task LSTM model for activity classification and intensity estimation.
[146]	CNN	SU	IAR	NA	Custom, SA6	94.2 92.5	A new CNN network for real-time AR
[138]	CNN+ LSTM+ Attention	SS	IAR	-k-means clustering, labeling flow, glimpse layer.	SA6, SA7, SA12	94.05 83.42 81.32	A pattern-balanced co-training for extracting the latent features from imbalanced and limited labeled AR data
[147]	CNN	SU	IAR	Sliding window, imaging	SA22	99.23	1) An unobtrusive DL-based AR for monitoring elderly peoples; 2) a method for encoding binary sensor logs into images.
[22]	CNN	SU	IAR	linear interpolation normalization Sliding window	SA4	89.6	Interpretable CNN for better classification based important sensor signals using spatially sparse convolutions
[148]	PSDRNN (LSTM)	SU	IAR	sliding window, PSD features, time and frequency features.	SA3	94.66	1) Extract PSD features from linear accelerations and tri-axle accelerations; 2) employ LSTM to learn these features.
[141]	TCN, LSTM	SU	GAR	trajectory segmentation, augmentation, sliding window	Custom	99.49, 99.39	1) A LIDAR point is clustered with the DBSCAN to carry out personal and object classification based on geometric features; 2) Employ TCN /LSTM to track multi-person concurrently and aggregate the corresponding trajectories from a 2D LiDAR.
[140]	LSTM	US	IAR	Segmentation, compression, L1-norm, low-pass elliptical filter, AHRS filter	Custom	95.15	1) A new unsupervised online learning scheme to avoid the constraints on the number of class constraints. 2) the Hk-mC algorithm is employed to label raw signals and build a hierarchical classification automatically.
[149]	ResNet	SU	IAR	GADF Imaging, GASF Imaging	SA7, SA8	9676, 99.2	1) A method for encoding sensory data into image representations; 2) A ResNet designed for extracting activity features from the encoded images.
[150]	RNN	SU	IAR	Statistical features, Augmentation, KPCA finetuning.	SA7, SA16,	99.3 99.1	1) An RNN based on effective features from the various wearable sensor; 2) employ KPCA to project features into a nonlinear space.
[151]	CNN+ Attention	SU	IAR	Sliding window Segmentation	SA1	96.4	1) A multi-head CNN introduced for robust feature extraction; 2) Attention mechanism designed for extra valuable feature selection.

[139]	LSTM + DQN	SS	IAR	Cleaning incomplete, incorrect pieces auto-segmentation	SA3, SA5	97.0	1) A semi-supervised framework is designed with LSTM for classifying sensory data; 2) A smart auto-labeling method based on DQN with a distance-based reward rule.
[152]	LSTM+ TE	SS	IAR	Augmentation, time, and frequency statistical features	SA6	97.11	1) A temporal ensembling of LSTM is designed for AR from huge and low-cost labeled and unlabeled sensor data.

LS=" Learning Strategy", SU=" Supervised", US=" Unsupervised", SS="Semi-supervised", A= "Accuracy"

3) RF-based activity recognition

The adoption of radio signals for recording human activities offers distinctive advantages because of the ubiquitous availability, which alleviates the privacy concerns incurred by vision-based and sensor-based approaches. The RF signals are also affected by the obstacles (walls) or by the darkness, which makes them perfect for modeling human activities, especially in indoor environments. Indeed, RF-based AR has come to be an enthusiastic research area in the latest years. A variety of commercial RF solutions have been presented for sensing, detecting, and recognizing different kinds of human activities. Table x overviews the recent research on DL for AR from different RF signals.

It can be observed that all of the reviewed DL studies emphasize recognizing the single-user activities (IAR level), making the GAR and HAR remain unexplored areas. This might be caused by the complexity of capturing the fluctuations of multi-user activities in radio signals. Similar to sensor-based approaches, the RF-based approaches are trained in a supervised manner, while semi-supervised training is employed using GAN designed for fall and gesture recognition [153]. Learning from unlabeled RF data can be improved by further exploration of the semi-supervised and unsupervised models. Besides, most of the reviewed studies experiment and evaluate their model on their custom datasets despite which necessitate reproducing their results on public data to understand the advantages and weakness of their models. Moreover, among the different kinds of RF data, the CSI has gained wide adoption in recognizing human activities, where most of them are collected using Intel 5300 NIC.

Table X. An overview of deep Learning studies for RF-based activity recognition

Ref	Model	LS	Level	Preparation	Dataset	Signal	A (%)	Contributions
[154]	LSTM	SU	IAR	Denoising Augmentation	Custom (Atheros)	CSI	97.8	An LSTM is employed for extracting features and for recognizing activities from differential CSI data.
[155]	CNN	SU	IAR	Butterworth filter, Segmentation	Custom (Intel 5300)	CSI	97.6	A two-stream CNN framework for learning the spatial and temporal patterns from CSI clips.
[156]	ABLSTM	SU	IAR	sliding window segmentation	Custom (Intel 5300), RA15	CSI	97.3, 97.0	1) leverage the BLSTM for extracting the sequential features from CSI streams in both forward and backward directions. 2) employ the attention layer to capture the significance of features learned by the BLSTM.
[157]	CNN	SU	IAR	Interpolation, Butterworth filter, Phase calibration, PCA, DTW	Custom (Intel 5300)	CSI	97-99.23	A three-phase framework to recognize multi-individual activities, where the layout of each phase depends on the size of the dataset.
[158]	LSTM	SU	IAR	sliding window segmentation, transformation, data denoising, PCA, DTW	Custom (Intel 5300)	CSI	90.64	1) An LSTM is employed to recognize handwriting actions from CSI data. 2) An CSI-Ratio model and relevant activity factor can be introduced to extract the segments of handwriting activity.
[159]	CNN+ Bi-LSTM	SU	IAR	data denoising, data segmentation, Snippet Action Acquisition	Custom (Intel 5300)	CSI	97.0	1) A DL framework employs CNN subnetwork to learn spatial dependencies, while Bi-LSTM is used to capture temporal information simultaneously. 2) A transfer learning scheme is presented to finetune the model performance in a new environment.
[160]	CNN+ LSTM	SU	IAR	Butterworth filter, PCA, STFT	Custom (Intel 5300)	CSI	96	A pattern-balanced co-training for extracting the latent features from imbalanced and limited labeled AR data

[161]	CNN	SU	IAR	phase processing, KL-divergence segmentation, PCA, STFT	Custom (M6e UHF)	RFID	95	A DL model for activity recognition from RFID spectrograms, which entail three modules namely CNN for feature encoding, dense module for activity classification, and a domain discriminator.
[153]	GAN	SS	IAR	NA	RA7, RA9	CSI	84.17, 84.09	1) A GAN-based model to solve the performance declination of leave-out validation for AR. 2) Generator, loss term, and Manifold regularization are used to learn from unlabeled data.
[162]	CNN	SU	IAR	sliding window, state inference	Custom (Intel 5300)	CSI	>91	1) A DL-based activity segmentation to lessen the reliance on knowledge and improve AR performance for blended activities. 2) A feedback system relating the segmentation with the classification method by jointly training them.
[163]	CNN+ Bi-LSTM	SU	IAR	Butterworth filter, PCA, STFT, Augmentation, Deformation	Custom (Intel 5300)	CSI	90.0	1) A data augmentation technique for synthesizing the CSI spectrogram to mitigate the impact of motion inconsistency and subjectivity issues; 2) A DL framework dedicated to learning from tiny CSI datasets and alleviating the overfitting problems.

LS=" Learning Strategy", SU=" Supervised", US=" Unsupervised", SS="Semi-supervised", A= "Accuracy"

VI. EMERGING MATTERS AND FUTURE DIRECTIONS

This section mainly discusses the most interesting research directions for both device-dependent and device-independent approaches in indoor IoT applications. In this regard, Table XI tabulates the key challenges facing the development of intelligent indoor IoT applications and the possible solutions based on the recent studies of intelligent IoT research.

Table XI: Challenges and possible solutions for different IoT applications in indoor environments.

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 tracks 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. - 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]	✓	✓

1) Transfer learning

Nowadays, DL approaches have gained control of developing intelligent IoT like the present tendency in the computer vision community. Nevertheless, training the new DL techniques from scratch remains a challenging mission to develop reliable applications. Consequently, the implementation of the DL approaches depending on priorly pre-trained architectures is a respectable strategy as these architectures have already experienced the underlying data representations. It is interesting to investigate the concept of transfer learning for some indoor trending applications using visual or sensory data streams.

2) Explainable Deep Learning

In recent days, the interpretability of visioning models has been becoming an extremely important research topic. Nevertheless, few research studies have been performed on explainable video recognition models. As clarified in [85,86], only some keyframes are critical for recognizing indoor activities, gestures, or indoor positions in a sequence of video frames taken out from the targeted video. Besides, the indoor activities/gestures vary in the corresponding temporal features. It is conceivable to recognize some activities/gestures utilizing the captured frames at the start or end of the video. The interpretability of complex activities/gestures depending on the keyframes is a virtuous research area to respond to the following questions, such as the arrangement of frames in the temporal domain? What is the contribution of the keyframes in the classification decision? and can these frames be designated for training the DL approach rapidly without impacting the efficiency of the Indoor applications? Such kind of understanding could help researchers to develop more effective IoT applications in the indoor environment.

3) Multimodal data

Indoor environments often contain multimodal data, including daily audio, visual, textual, and signal, generated and received by humans to communicate with the surrounding environment. As an example, reading allows the rebuilding of the consistent portion of the individual's visual intellect. Thus, it is advantageous to exploit multimodal information to understand complex indoor activities because multimodal data encompass amusing semantic information [87]. Modeling this kind of data enables acquiring the long-standing temporal interdependency among entities from the multimodal data since it could be thought-provoking to straightforwardly learn from the multimodal data [88]. This long-standing temporal interdependency could show the consecutive order of indoor activities/gestures/positions throughout a lengthy sequence similar to how the human brain performs. Once a person recalls somewhat, one item induces the following item from the lengthy main sequence, comparable to enduring video. Moreover, the interaction among various entities is also significant to comprehend temporal interdependency. As an example, predetermined interactions between objects occur in a specific activity following certain conditions. Therefore, the indoor DL-based IoT application should consider the multimodal information about humans to enable reliable performance, especially in applications that depend on long-duration data.

4) The physical aspect of humans.

Nowadays, there is an increasing curiosity in investigating the bodily facets of human actions, like detailed and specific activities/gestures. For instance, authors in [89] presented a 20BN-something-something as a HAR dataset to inspire the researcher to explore human-object relations. This dataset covers class patterns or documentary descriptions, such as "set an entity near to an object" to represent the interaction between human and object or between two objects. Such kind of data enables developing an indoor IoT application that considers the bodily facets of human movements/actions counting the interactions between human and object and the spatial relationships. Although many statistics are detected with the Closed-circuit television (CCTV) videos, some bodily facets, like power, speed, movement style, and rushing, are difficult to capture. Thus, it is vitally important to develop an IoT benchmark that contains such kind of information.

5) Learning actions without labels

In an attempt to enlarge the size of the indoor dataset used to train the DL model from any application domains, the physical annotation of data samples is labor-intensive, ineffective and expensive. Even though the automated annotation exploiting the search engines and video subtitles is realizable in some fields, it still requires manual confirmation. Crowdsourcing [104] has been suggested as a healthier solution. Nevertheless, it is challenging due to the label multiplicity issue, resulting in an inappropriate outcome. Therefore, the research community requires introducing improved and powerful learning techniques that inevitably manipulate the unannotated generated indoor data [140].

VII. CONCLUSION

A thorough survey of the recent cutting-edge deep learning approaches accompanied by their positives and negatives cons for device-dependent and device-independent has been introduced in this work. These approaches have become particularly prominent in the latest years due to their potential integration in different IoT applications of smart indoor environments, including positioning and activity recognition applications. The comprehensive explanations, analyses, and insights of the corresponding aspects assist researchers in enriching their knowledge in the era of human-centered IoT applications with indoor environments.

Several viewpoints have been considered in discussing the current studies, including DL architecture, accuracy, application, system configuration, used data, sensors, samples. We put emphasis on the latest advancement in both device-dependent and device-independent IoT applications. We provided a novel taxonomy for smart indoor DL approaches from a data modality perspective and/or application domain perspective. The characteristics, advantages, and flaws of contemporary DL approaches employed in indoor IoT applications were also studied. Furthermore, this survey study discusses the most interesting research issues facing Indoor IoT applications and presents some potential solutions for these issues.

Above and beyond the DL applications in different indoor environments, various challenging issues are fruitful for future investigation, such as system design, multi-activity tracking, action forecasting, and time sensitivity. This study is likely to inspire more research in a wide variety of human-centered IoT applications in indoor environments.

REFERENCES

[1] J. Bai, S. Lian, Z. Liu, K. Wang, and Di. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Trans. Consum. Electron.*, 2017, doi: 10.1109/TCE.2017.014980.

[2] Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi Vision: Sensing, Recognition, and Detection with Commodity MIMO-OFDM WiFi," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2989426.

[3] G. Oguntala, Y. F. Hu, A. A. S. Alabdullah, R. Abd-Alhameed, M. Ali, and D. Luong, "Passive RFID Module with LSTM Recurrent Neural Network Activity Classification Algorithm for Ambient Assisted Living," *IEEE Internet Things J.*, 2021, doi:

10.1109/JIOT.2021.3051247.
[4] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Communications Surveys and Tutorials*. 2018, doi: 10.1109/COMST.2018.2844341.
[5] R. Zhang, X. Jing, S. Wu, C. Jiang, J. Mu, and F. Richard Yu, "Device-Free Wireless Sensing for Human Detection: The Deep Learning Perspective," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3024234.
[6] J. Zhang and D. Tao, "Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.3039359.
[7] C. Yin, S. Zhang, J. Wang, and N. N. Xiong, "Anomaly Detection Based on Convolutional Recurrent Autoencoder for IoT Time Series," *IEEE Trans. Syst. Man, Cybern. Syst.*, 2020, doi: 10.1109/tsmc.2020.2968516.
[8] F. Alam, N. Faulkner, and B. Parr, "Device-Free Localization: A Review of Non-RF Techniques for Unobtrusive Indoor Positioning," *IEEE Internet of Things Journal*. 2021, doi: 10.1109/JIOT.2020.3030174.
[9] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN Combined with LSTM on video kinematic data," *IEEE J. Biomed. Heal. Informatics*, 2019, doi: 10.1109/JBHI.2018.2808281.
[10] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, 2019, doi: 10.1016/j.inffus.2018.06.002.
[11] Y. Ye, Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, 2020, doi: 10.1109/TSG.2020.2976771.
[12] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, 2020, doi: 10.1016/j.patcog.2020.107561.
[13] Z. Hussain, Q. Z. Sheng, and W. E. Zhang, "A review and categorization of techniques on device-free human activity recognition," *Journal of Network and Computer Applications*. 2020, doi: 10.1016/j.jnca.2020.102738.
[14] Q. Li, R. Gravina, Y. Li, S. H. Alsamhi, F. Sun, and G. Fortino, "Multi-user activity recognition: Challenges and opportunities," *Inf. Fusion*, 2020, doi: 10.1016/j.inffus.2020.06.004.
[15] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen, "Sensor-Based Continuous Authentication of Smartphones' Users Using Behavioral Biometrics: A Contemporary Survey," *IEEE Internet of Things Journal*. 2021, doi: 10.1109/JIOT.2020.3020076.
[16] X. Guo, N. Ansari, F. Hu, Y. Shao, N. R. Elikplim, and L. Li, "A survey on fusion-based indoor positioning," *IEEE Commun. Surv. Tutorials*, 2020, doi: 10.1109/COMST.2019.2951036.
[17] X. Zhu, W. Qu, T. Qiu, L. Zhao, M. Atiquzzaman, and D. O. Wu, "Indoor Intelligent Fingerprint-Based Localization: Principles, Approaches and Challenges," *IEEE Commun. Surv. Tutorials*, 2020, doi: 10.1109/COMST.2020.3014304.
[18] S. Deep, X. Zheng, C. Karmakar, D. Yu, L. G. C. Hamey, and J. Jin, "A Survey on Anomalous Behavior Detection for Elderly Care Using Dense-Sensing Networks," *IEEE Commun. Surv. Tutorials*, 2020, doi: 10.1109/COMST.2019.2948204.
[19] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Zhu, "Deep Learning for Radio-based Human Sensing: Recent Advances and Future Directions," *IEEE Commun. Surv. Tutorials*, 2021, doi: 10.1109/COMST.2021.3058333.
[20] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless Sensing for Human Activity: A Survey," *IEEE Commun. Surv. Tutorials*, 2020, doi: 10.1109/COMST.2019.2934489.
[21] H. F. Thariq Ahmed, H. Ahmad, and A. CV, "Device free human gesture recognition using Wi-Fi CSI: A survey," *Eng. Appl. Artif. Intell.*, 2020, doi: 10.1016/j.engappai.2019.103281.
[22] E. Kim, "Interpretable and Accurate Convolutional Neural Networks for Human Activity Recognition," *IEEE Trans. Ind. Informatics*, 2020, doi: 10.1109/TII.2020.2972628.
[23] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A Survey on Behavior Recognition Using WiFi Channel State Information," *IEEE Commun. Mag.*, 2017, doi: 10.1109/MCOM.2017.1700082.
[24] L. Zheng, B. J. Hu, J. Qiu, and M. Cui, "A Deep-Learning-Based Self-Calibration Time-Reversal Fingerprinting Localization Approach on Wi-Fi Platform," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2981723.
[25] F. Gu, K. Khoshelham, C. Yu, and J. Shang, "Accurate Step Length Estimation for Pedestrian Dead Reckoning Localization Using Stacked Autoencoders," *IEEE Trans. Instrum. Meas.*, 2019, doi: 10.1109/TIM.2018.2871808.
[26] M. Kim, D. Han, and J. K. Rhee, "Multiview Variational Deep Learning with Application to Practical Indoor Localization," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2021.3063512.
[27] I. Sohn, "Deep belief network based intrusion detection techniques: A survey," *Expert Systems with Applications*. 2021, doi: 10.1016/j.eswa.2020.114170.
[28] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A Cost-Sensitive Deep Belief Network for Imbalanced Classification," *IEEE Trans. Neural Networks Learn. Syst.*, 2019, doi: 10.1109/TNNLS.2018.2832648.
[29] Z. Wang, Q. She, and T. Ward, "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," *ACM Comput. Surv.*, 2021, doi: 10.1145/3439723.
[30] N. Gao *et al.*, "Generative adversarial networks for spatio-temporal data: A survey," *arXiv*. 2020.
[31] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, 2019, doi: 10.1016/j.media.2019.03.009.
[32] T. Guo, C. Xu, S. He, B. Shi, C. Xu, and D. Tao, "Robust Student Network Learning," *IEEE Trans. Neural Networks Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2019.2929114.
[33] X. Li, L. Yu, H. Chen, C. W. Fu, L. Xing, and P. A. Heng, "Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation," *IEEE Trans. Neural Networks Learn. Syst.*, 2021, doi: 10.1109/TNNLS.2020.2995319.
[34] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017.
[35] D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019.
[36] K. Sohn *et al.*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv*. 2020.
[37] H. Zhang, Z. Hu, W. Qin, M. Xu, and M. Wang, "Adversarial co-distillation learning for image recognition," *Pattern Recognit.*, 2021, doi: 10.1016/j.patcog.2020.107659.
[38] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," 2009, doi: 10.1109/CVPRW.2009.5206557.
[39] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre Thomas, "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, 2013.
[40] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Dec. 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.
[41] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," 2007.
[42] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," 2012, doi: 10.1109/CVPRW.2012.6239234.
[43] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Rob. Res.*, 2013, doi: 10.1177/0278364913478446.
[44] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive Multimodal Human Action Database," 2013, doi: 10.1109/WACV.2013.6474999.
[45] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," 2014, doi: 10.1109/CVPR.2014.223.
[46] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," 2015, doi: 10.1109/ICIP.2015.7350781.
[47] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," 2016, doi: 10.1109/CVPR.2016.115.
[48] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2019.2916873.
[49] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," 2015, doi: 10.1109/CVPR.2015.7298698.
[50] P. Weinzaepfel, X. Martin, and C. Schmid, "Human Action Localization with Sparse Spatial Supervision," May 2016, [Online]. Available: http://arxiv.org/abs/1605.05197.
[51] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and Observer: Joint Modeling of First and Third-Person Videos," 2018, doi: 10.1109/CVPR.2018.00772.
[52] R. Goyal *et al.*, "The 'Something Something' Video Database for Learning and Evaluating Visual Common Sense," 2017, doi: 10.1109/ICCV.2017.622.
[53] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos," *Int. J. Comput. Vis.*, 2018, doi: 10.1007/s11263-017-1013-y.
[54] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A Short Note on the Kinetics-700 Human Action Dataset," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.06987.
[55] C. Gu *et al.*, "AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions," 2018, doi: 10.1109/CVPR.2018.00633.
[56] M. Monfort *et al.*, "Moments in Time Dataset: One Million Videos for Event Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2019.2901464.
[57] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," 2019, doi: 10.1109/ICCV.2019.00876.
[58] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Trans. Affect. Comput.*, 2015, doi: 10.1109/TAFFC.2015.2397456.
[59] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," 2009, doi: 10.1109/ICCV.2009.5459361.
[60] S. Blunsden and R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification," *Ann. BMVA*, 2010.
[61] W. R. Ko, M. Jang, J. Lee, and J. Kim, "AIR-Act2Act: Human–human interaction dataset for teaching non-verbal social behaviors to robots," *Int. J. Rob. Res.*, 2021, doi: 10.1177/0278364921990671.
[62] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," 2009, doi: 10.1109/ICCVW.2009.5457461.
[63] R. Chavarriaga *et al.*, "The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, 2013, doi: 10.1016/j.patrec.2012.12.014.
[64] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," 2013.
[65] O. Banos *et al.*, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomed. Eng. Online*, 2015, doi: 10.1186/1475-925X-14-S2-S6.
[66] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newsl.*, 2011, doi: 10.1145/1964897.1964918.
[67] G. M. Weiss and J. W. Lockhart, "The impact of personalization on smartphone-based activity recognition," 2012.
[68] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, 2017, doi: 10.3390/app7101101.
[69] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," 2016, doi: 10.1109/PERCOM.2016.7456521.
[70] A. Stisen *et al.*, "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," 2015, doi: 10.1145/2809695.2809718.
[71] T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, "Daily activity recognition based on DNN using environmental sound and acceleration signals," 2015, doi: 10.1109/EUSIPCO.2015.7362796.
[72] B. Barshan and M. C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *Comput. J.*, 2013, doi: 10.1093/comjnl/bxt075.
[73] N. Sikder and A.-A. Nahid, "KU-HAR: An open dataset for heterogeneous human activity recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 46–54, Jun. 2021, doi: 10.1016/j.patrec.2021.02.024.
[74] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," 2012, doi: 10.1109/ISWC.2012.13.
[75] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, Multi-Sensor Based Classification of Daily Life Activities: Comparison with State-of-the-Art Algorithms Using a Benchmark Dataset," *PLoS One*, 2013, doi: 10.1371/journal.pone.0075196.
[76] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human

pose from sparse inertial measurements in real time," 2018, doi: 10.1145/3272127.3275108.
[77] H. Leutheuser, S. Doelfel, D. Schuldhaus, S. Reinfelder, and B. M. Eskofier, "Performance comparison of two step segmentation algorithms using different step activities," 2014, doi: 10.1109/BSN.2014.37.
[78] W. Ugulino, D. Cardador, K. Vega, E. Velloso, R. Milidiú, and H. Fuks, "Wearable computing: Accelerometers' data classification of body postures and movements," 2012, doi: 10.1007/978-3-642-34459-6_6.
[79] R. Wang *et al.*, "Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," 2014, doi: 10.1145/2632048.2632054.
[80] S. Rossi, R. Capasso, G. Acampora, and M. Staffa, "A Multimodal Deep Learning Network for Group Activity Recognition," 2018, doi: 10.1109/IJCNN.2018.8489309.
[81] D. Gordon, M. Wirz, D. Roggen, G. Tröster, and M. Beigl, "Group affiliation detection using model divergence for wearable devices," 2014, doi: 10.1145/2634317.2634319.
[82] H. Alemdar, H. Ertan, O. D. Incel, and C. Ersoy, "ARAS human activity datasets in multiple homes with multiple residents," 2013, doi: 10.4108/icst.pervasivehealth.2013.252120.
[83] "CASAS Smart Home Project." http://casas.wsu.edu/datasets/ (accessed Mar. 25, 2021).
[84] R. Alazrai, A. Awad, B. Alsaify, M. Hababeh, and M. I. Daoud, "A dataset for Wi-Fi-based human-to-human interaction recognition," *Data Br.*, 2020, doi: 10.1016/j.dib.2020.105668.
[85] L. Guo *et al.*, "Wiar: A public dataset for wifi-based activity recognition," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2947024.
[86] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "CrossSense: Towards cross-site and large-scale WiFi sensing," 2018, doi: 10.1145/3241539.3241570.
[87] P. Hillyard *et al.*, "Experience: Cross-technology radio respiratory monitoring performance study," 2018, doi: 10.1145/3241539.3241560.
[88] J. K. Brinke and N. Meratnia, "Dataset: Channel state information for different activities, participants and days," 2019, doi: 10.1145/3359427.3361913.
[89] Y. Zheng *et al.*, "Zero-effort cross-domain gesture recognition with Wi-Fi," 2019, doi: 10.1145/3307334.3326081.
[90] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using WiFi," 2017, doi: 10.1145/3081333.3081340.
[91] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign Language Recognition Using WiFi," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2018, doi: 10.1145/3191755.
[92] M. A. A. Haseeb and R. Parasuraman, "Wisture: RNN-based learning of wireless signals for gesture recognition in unmodified smartphones," *arXiv*. 2017.
[93] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDeFi: Ubiquitous Fall Detection using Commodity Wi-Fi Devices," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2018, doi: 10.1145/3161183.
[94] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," 2019, doi: 10.1145/3349624.3356768.
[95] F. Wang, J. Han, S. Zhang, X. He, and D. Huang, "CSI-Net: Unified Human Body Characterization and Pose Recognition," *arXiv*. 2018.
[96] I. Sobron, J. Del Ser, I. Eizmendi, and M. Velez, "Device-Free People Counting in IoT Environments: New Insights, Results, and Open Challenges," *IEEE Internet Things J.*, 2018, doi: 10.1109/JIOT.2018.2806990.
[97] Z. Meng *et al.*, "Gait Recognition for Co-Existing Multiple People Using Millimeter Wave Sensing," vol. 34, no. 01, pp. 849–856, 2020, doi: https://ojs.aaai.org/index.php/AAAI/article/view/5430.
[98] G. Jekabsons and V. Zuravlyovs, "Refining Wi-Fi Based Indoor Positioning," *Aict2010 - Appl. Inf. Commun. Technol. Proc. 4Th Int. Sci. Conf.*, 2010.
[99] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with WiFi fingerprints," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2923743.
[100] J. Torres-Sospedra *et al.*, "The smartphone-based offline indoor location competition at IPIN 2016: Analysis and future work," *Sensors (Switzerland)*, 2017, doi: 10.3390/s17030557.
[101] J. Torres-Sospedra *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," 2014, doi: 10.1109/IPIN.2014.7275492.
[102] J. Torres-Sospedra, D. Rambla, R. Montoliu, O. Belmonte, and J. Huerta, "UJIIndoorLoc-Mag: A new database for magnetic field-based localization problems," 2015, doi: 10.1109/IPIN.2015.7346763.
[103] R. Montoliu, E. Sansano, J. Torres-Sospedra, and O. Belmonte, "IndoorLoc platform: A public repository for comparing and evaluating indoor positioning systems," 2017, doi: 10.1109/IPIN.2017.8115940.
[104] E. S. Lohan, J. Torres-Sospedra, H. Leppäkoski, P. Richter, Z. Peng, and J. Huerta, "Wi-Fi crowdsourced fingerprinting dataset for indoor positioning," *Data*, 2017, doi: 10.3390/data2040032.
[105] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi," 2017, doi: 10.1145/3084041.3084067.
[106] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2.0: Passive human tracking with a single Wi-Fi link," 2018, doi: 10.1145/3210240.3210314.
[107] Y. Zhao, J. Xu, J. Wu, J. Hao, and H. Qian, "Enhancing Camera-Based Multimodal Indoor Localization with Device-Free Movement Measurement Using WiFi," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2019.2948605.
[108] I. Ha, H. Kim, S. Park, and H. Kim, "Image retrieval using BIM and features from pretrained VGG network for indoor localization," *Build. Environ.*, 2018, doi: 10.1016/j.buildenv.2018.05.026.
[109] P. Chhikara, R. Tekchandani, N. Kumar, V. Chamola, and M. Guizani, "DCNN-GA: A Deep Neural Net Architecture for Navigation of UAV in Indoor Environment," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3027095.
[110] N. Lee, S. Ahn, and D. Han, "AMID: Accurate magnetic indoor localization using deep learning," *Sensors (Switzerland)*, 2018, doi: 10.3390/s18051598.
[111] X. Wang, Z. Yu, and S. Mao, "Indoor Localization Using Smartphone Magnetic and Light Sensors: a Deep LSTM Approach," *Mob. Networks Appl.*, 2020, doi: 10.1007/s11036-019-01302-x.
[112] Y. Zhu, H. Luo, F. Zhao, and R. Chen, "Indoor/Outdoor Switching Detection Using Multisensor DenseNet and LSTM," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1544–1556, Feb. 2021, doi: 10.1109/JIOT.2020.3013853.
[113] X. Wang, X. Wang, and S. Mao, "Deep Convolutional Neural Networks for Indoor Localization with CSI Images," *IEEE Trans. Netw. Sci. Eng.*, 2020, doi: 10.1109/TNSE.2018.2871165.
[114] L. Zhang *et al.*, "WiFi-Based Indoor Robot Positioning Using Deep Fuzzy Forests," *IEEE Internet Things J.*, 2020, doi:

10.1109/JIOT.2020.2986685.
[115] Y. Xue, W. Su, H. Wang, D. Yang, and Y. Jiang, "DeepTAL: Deep Learning for TDOA-Based Asynchronous Localization Security with Measurement Error and Missing Data," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2937975.
[116] J. He and H. C. So, "A Hybrid TDOA-Fingerprinting-Based Localization System for LTE Network," *IEEE Sens. J.*, 2020, doi: 10.1109/JSEN.2020.3004179.
[117] W. Wang, P. Bai, Y. Zhou, X. Liang, and Y. Wang, "Optimal Configuration Analysis of AOA Localization and Optimal Heading Angles Generation Method for UAV Swarms," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2918299.
[118] A. Khan, S. Wang, and Z. Zhu, "Angle-of-Arrival Estimation Using an Adaptive Machine Learning Framework," *IEEE Commun. Lett.*, 2019, doi: 10.1109/LCOMM.2018.2884464.
[119] Y. Zheng, M. Sheng, J. Liu, and J. Li, "Exploiting aoa estimation accuracy for indoor localization: A weighted aoa-based approach," *IEEE Wirel. Commun. Lett.*, 2018, doi: 10.1109/LWC.2018.2853745.
[120] X. Wang, X. Wang, S. Mao, J. Zhang, S. C. G. Periaswamy, and J. Patton, "Indoor Radio Map Construction and Localization with Deep Gaussian Processes," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2996564.
[121] M. Chen *et al.*, "MoLoc: Unsupervised Fingerprint Roaming for Device-Free Indoor Localization in a Mobile Ship Environment," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.3004240.
[122] Y. Li, X. Hu, Y. Zhuang, Z. Gao, P. Zhang, and N. El-Sheimy, "Deep Reinforcement Learning (DRL): Another Perspective for Unsupervised Wireless Localization," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2019.2957778.
[123] X. Wang, X. Wang, and S. Mao, "Indoor Fingerprinting with Bimodal CSI Tensors: A Deep Residual Sharing Learning Approach," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3026608.
[124] R. Wang, H. Luo, Q. Wang, Z. Li, F. Zhao, and J. Huang, "A Spatial-Temporal Positioning Algorithm Using Residual Network and LSTM," *IEEE Trans. Instrum. Meas.*, 2020, doi: 10.1109/TIM.2020.2998645.
[125] Q. Wang *et al.*, "Pedestrian Dead Reckoning Based on Walking Pattern Recognition and Online Magnetic Fingerprint Trajectory Calibration," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3016146.
[126] D. Chen, S. Yongchareon, E. M. K. Lai, J. Yu, and Q. Z. Sheng, "Hybrid Fuzzy C-means CPD-based Segmentation for Improving Sensor-based Multi-resident Activity Recognition," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2021.3051574.
[127] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic Graph Neural Networks for 3D Skeleton-based Human Action Recognition and Motion Prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021, doi: 10.1109/TPAMI.2021.3053765.
[128] C. Dhiman and D. K. Vishwakarma, "View-Invariant Deep Architecture for Human Action Recognition Using Two-Stream Motion and Shape Temporal Dynamics," *IEEE Trans. Image Process.*, 2020, doi: 10.1109/TIP.2020.2965299.
[129] T. Hussain *et al.*, "Multi-View Summarization and Activity Recognition Meet Edge Computing in IoT Environments," *IEEE Internet Things J.*, 2020, doi: 10.1109/jiot.2020.3027483.
[130] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi, "SGM-Net: Skeleton-guided multimodal network for action recognition," *Pattern Recognit.*, 2020, doi: 10.1016/j.patcog.2020.107356.
[131] F. Meng, H. Liu, Y. Liang, J. Tu, and M. Liu, "Sample Fusion Network: An End-to-End Data Augmentation Network for Skeleton-Based Human Action Recognition," *IEEE Trans. Image Process.*, 2019, doi: 10.1109/TIP.2019.2913544.
[132] X. Shu, J. Tang, G. J. Qi, W. Liu, and J. Yang, "Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021, doi: 10.1109/TPAMI.2019.2942030.
[133] X. Shu, L. Zhang, Y. Sun, and J. Tang, "Host-Parasite: Graph LSTM-in-LSTM for Group Activity Recognition," *IEEE Trans. Neural Networks Learn. Syst.*, 2021, doi: 10.1109/TNNLS.2020.2978942.
[134] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence Constrained Graph LSTM for Group Activity Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, doi: 10.1109/tpami.2019.2928540.
[135] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou, "Learning Semantics-Preserving Attention and Contextual Interaction for Group Activity Recognition," *IEEE Trans. Image Process.*, 2019, doi: 10.1109/tip.2019.2914577.
[136] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "HiGCIN: Hierarchical Graph-based Cross Inference Network for Group Activity Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/tpami.2020.3034233.
[137] L. Pei *et al.*, "MARS: Mixed Virtual and Real Wearable Sensors for Human Activity Recognition with Multi-Domain Deep Learning Model," *arXiv*. 2020, doi: 10.1109/jiot.2021.3055859.
[138] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A Semisupervised Recurrent Convolutional Attention Model for Human Activity Recognition," *IEEE Trans. Neural Networks Learn. Syst.*, 2020, doi: 10.1109/TNNLS.2019.2927224.
[139] X. Zhou, W. Liang, K. I. K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2985082.
[140] W. Qi, H. Su, and A. Aliverti, "A Smartphone-Based Adaptive Recognition and Real-Time Monitoring System for Human Activities," *IEEE Trans. Human-Machine Syst.*, 2020, doi: 10.1109/THMS.2020.2984181.
[141] F. Luo, S. Poslad, and E. Bodanese, "Temporal Convolutional Networks for Multiperson Activity Recognition Using a 2-D LIDAR," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7432–7442, Aug. 2020, doi: 10.1109/JIOT.2020.2984544.
[142] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, "TSE-CNN: A Two-Stage End-to-End CNN for Human Activity Recognition," *IEEE J. Biomed. Heal. Informatics*, 2020, doi: 10.1109/JBHI.2019.2909688.
[143] N. Rashid, M. Dautta, P. Tseng, and M. A. Al Faruque, "HEAR: Fog-Enabled Energy-Aware Online Human Eating Activity Recognition," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3008842.
[144] T. Huynh-The, C. H. Hua, N. A. Tu, and D. S. Kim, "Physical Activity Recognition with Statistical-Deep Fusion Model Using Multiple Sensory Data for Smart Health," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3013272.
[145] O. Barut, L. Zhou, and Y. Luo, "Multi-task LSTM Model for Human Activity Recognition and Intensity Estimation Using Wearable Sensor Data," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8760–8768, Sep. 2020, doi: 10.1109/JIOT.2020.2996578.
[146] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, "IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment," *IEEE Internet Things J.*, 2019, doi: 10.1109/JIOT.2019.2920283.
[147] M. Gochoo, T. H. Tan, S. H. Liu, F. R. Jean, F. S. Alnajjar, and S. C. Huang, "Unobtrusive Activity Recognition of Elderly People Living Alone Using Anonymous Binary Sensors and DCNN," *IEEE J. Biomed. Heal. Informatics*, 2019, doi: 10.1109/JBHI.2018.2833618.
[148] X. Li, Y. Wang, B. Zhang, and J. Ma, "PSDRNN: An Efficient and Effective HAR Scheme Based on Feature Extraction and Deep Learning," *IEEE Trans. Ind. Informatics*, 2020, doi: 10.1109/TII.2020.2968920.

[149] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K. K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Inf. Fusion*, 2020, doi: 10.1016/j.inffus.2019.06.014.
[150] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, "A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare," *Inf. Fusion*, 2020, doi: 10.1016/j.inffus.2019.08.004.
[151] H. Zhang, Z. Xiao, J. Wang, F. Li, and E. Szczerbicki, "A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multihead Convolutional Attention," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2019.2949715.
[152] Q. Zhu, Z. Chen, and Y. C. Soh, "A Novel Semisupervised Deep Learning Method for Human Activity Recognition," *IEEE Trans. Ind. Informatics*, vol. 15, no. 7, pp. 3821–3830, Jul. 2019, doi: 10.1109/TII.2018.2889315.
[153] C. Xiao, D. Han, Y. Ma, and Z. Qin, "CsiGAN: Robust Channel State Information-Based Activity Recognition With GANs," *IEEE Internet Things J.*, 2019, doi: 10.1109/JIOT.2019.2936580.
[154] P. Khan, B. S. K. Reddy, A. Pandey, S. Kumar, and M. Youssef, "Differential Channel-State-Information-Based Human Activity Recognition in IoT Networks," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2997237.
[155] B. Sheng, Y. Fang, F. Xiao, and L. Sun, "An Accurate Device-Free Action Recognition System Using Two-Stream Network," *IEEE Trans. Veh. Technol.*, 2020, doi: 10.1109/TVT.2020.2993901.
[156] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM," *IEEE Trans. Mob. Comput.*, 2019, doi: 10.1109/TMC.2018.2878233.
[157] C. Feng, S. Arshad, S. Zhou, D. Cao, and Y. Liu, "Wi-Multi: A Three-Phase System for Multiple Human Activity Recognition with Commercial WiFi Devices," *IEEE Internet Things J.*, 2019, doi: 10.1109/JIOT.2019.2915989.
[158] Z. Guo, F. Xiao, B. Sheng, H. Fei, and S. Yu, "WiReader: Adaptive Air Handwriting Recognition Based on Commercial WiFi Signal," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2997053.
[159] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep Spatial-Temporal Model Based Cross-Scene Action Recognition Using Commodity WiFi," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2973272.
[160] F. Wang, W. Gong, and J. Liu, "On spatial diversity in wifi-based human activity recognition: A deep learning-based approach," *IEEE Internet Things J.*, 2019, doi: 10.1109/JIOT.2018.2871445.
[161] F. Wang, J. Liu, and W. Gong, "Multi-Adversarial In-Car Activity Recognition using RFIDs," *IEEE Trans. Mob. Comput.*, 2020, doi: 10.1109/tmc.2020.2977902.
[162] C. Xiao, Y. Lei, Y. Ma, F. Zhou, and Z. Qin, "DeepSeg: Deep Learning-based Activity Segmentation Framework for Activity Recognition using WiFi," *IEEE Internet Things J.*, 2020, doi: 10.1109/jiot.2020.3033173.
[163] J. Zhang *et al.*, "Data Augmentation and Dense-LSTM for Human Activity Recognition Using WiFi Signal," *IEEE Internet Things J.*, 2021, doi: 10.1109/JIOT.2020.3026732.