A Survey on Deep Learning for Human Activity Recognition

FUQIANG GU, Chongqing University, China
MU-HUAN CHUNG, MARK CHIGNELL, and SHAHROKH VALAEE,
University of Toronto, Canada
BAODING ZHOU, Shenzhen University, China
XUE LIU, McGill University, Canada

Human activity recognition is a key to a lot of applications such as healthcare and smart home. In this study, we provide a comprehensive survey on recent advances and challenges in human activity recognition (HAR) with deep learning. Although there are many surveys on HAR, they focused mainly on the taxonomy of HAR and reviewed the state-of-the-art HAR systems implemented with conventional machine learning methods. Recently, several works have also been done on reviewing studies that use deep models for HAR, whereas these works cover few deep models and their variants. There is still a need for a comprehensive and in-depth survey on HAR with recently developed deep learning methods.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing;**

Additional Key Words and Phrases: Machine learning, deep learning, activity recognition, mobile sensing, deep models

ACM Reference format:
Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. 2021. A Survey on Deep Learning for Human Activity Recognition. *ACM Comput. Surv.* 54, 8, Article 177 (October 2021), 34 pages.
https://doi.org/10.1145/3472290

177

Jianga Shang, China University of Geoscience (Wuhan), China, has been added as an author for this paper after it was accepted.
This study is jointly supported by the National Natural Science Foundation of China (No. 42174050 and No. 42171427), National Key Research and Development Program of China (No. 2016YFB0502200), Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011910), and Shenzhen Scientific Research and Development Funding Program (No. JCYJ20190808113603556).
Authors' addresses: F. Gu, College of Computer Science, Chongqing University, 174 Shazheng Street, Chongqing, 400044, China; email: gufq@cqu.edu.cn; M.-H. Chung and M. Chignell, Department of Mechanical and Industrial Engineering, University of Toronto, 40 St George Street, Toronto, ON, M5S 2E4, Canada; emails: {mhmchung, chignell}@mie.utoronto.ca; S. Valaee, Department of Electrical and Computer Engineering, University of Toronto, 40 St George Street, Toronto, ON, M5S 2E4, Canada; email: valaee@ece.utoronto.ca; B. Zhou (corresponding author), College of Civil and Transportation Engineering, Institute of Urban Smart Transportation & Safety Maintenance, Guangdong Key Laboratory of Urban Informatics, Shenzhen University, Shenzhen, Guangdong, 518052, China; email: bdzhou@szu.edu.cn; X. Liu, School of Computer Science, McGill University, 845 Sherbrooke St W, Montreal, Quebec, H3A 0G4, Canada; email: xueliu@cs.mcgill.ca.

0360-0300/2021/10-ART177 $15.00
https://doi.org/10.1145/3472290

1 INTRODUCTION

The knowledge of human activity is crucial for a lot of applications and services such as health monitoring, fitness, home automation, augmented reality, traffic scheduling and control, precise advertising, and security [110]. For example, the record of a person's daily activities can be used to calculate the calorie he has consumed on a day, which can further suggest proper diet for him to keep healthy and fit; detecting the fall activity of elderly people can be used to trigger emergency assistance to avoid causing severe accidents.

Human activity can be recognized by applying conventional machine learning methods. However, conventional machine learning methods for **Human Activity Recognition (HAR)** require to design and select relevant features. This process involves laborious human intervention and expert knowledge, and the designed and selected features might still achieve suboptimal performance. To relieve the burden of hand-engineering features, deep learning methods have been proposed in recent years [112, 207]. Deep learning methods are very useful for HAR and can benefit HAR from several aspects. First, it relieves the effort of manually designing features, which often requires expert knowledge. Second, it has shown better accuracy in HAR than conventional methods [69, 160, 260]. Third, it has the ability to learn from unlabeled data, which is important and useful for HAR, since it is unpractical to obtain a large amount of labeled activity data. Fourth, it has the powerful capability of learning useful features from raw data and can deal with activity-related data from different people, different device models, and varying device poses.

The relationship of deep learning, machine learning, and artificial intelligence is demonstrated in Figure 1. Deep learning is a subset of machine learning methods, and has multiple levels of representations. Deep learning networks are artificial neural networks with more than one hidden layer, and therefore deep learning networks are also known as deep neural networks. In Reference [40], the authors categorized deep learning models as deep networks for supervised learning, deep networks for unsupervised learning, and hybrid approaches. In this work, we adapt the categorization in Reference [40] and divide deep learning models into deep generative models, deep discriminative models, and deep hybrid models. Deep generative models aim to learn useful representations of data via unsupervised learning or to learn the joint probability distribution of data and their associated classes [184]. Popular generative models are **Restricted Boltzmann Machines (RBMs)** [86], autoencoders [216, 219], **Generative Adversarial Networks (GANs)** [65], and their variants. Discriminative models aim to learn the conditional probability distribution of classes on the data, in which the label information is available directly or indirectly [239]. Popular deep discriminative models are **Convolutional Neural Networks (CNNs)** [62, 106, 178], **Recurrent Neural Networks (RNNs)** [20, 186], and their variants. Deep hybrid models combine a generative model and a discriminative model where the outcome of the generative model is often used as the input to the discriminative model for classification or regression [52]. These models were originally proposed for processing images, video, speech, and audio, but they can also be applied to other domains such as activity recognition [150, 181, 244] and indoor localization [67, 71].

So far, there have been several surveys on HAR in the literature. Poppe [164] reviewed research works on vision-based human action recognition and discussed different image representation methods as well as action classification methods. Aggarwal and Ryoo [3] presented an approach-based taxonomy for HAR and discussed recognition methods, including space-time approaches, sequential approaches, and hierarchical approaches, for simple human actions and high-level activities (e.g., human–human interactions and human–object interactions, group activities). Chen et al. [34] surveyed various aspects of sensor-based activity recognition, mainly including data-driven and knowledge-driven methods for activity monitoring, modeling, and recognition. Incel et al. [93] provided a taxonomy of activity recognition on smartphones, introduced the process

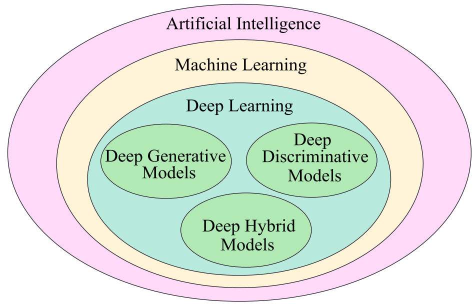


Fig. 1. The relationship of deep learning, machine learning, and artificial intelligence.

and challenges of HAR on phones, and reviewed HAR works based on location, motion, and other contextual information. Lara and Labrador [110] conducted a survey on HAR with wearable sensors, and discussed different types of activities, the design issues and recognition methods of HAR. Specifically, they evaluated 28 systems in terms of recognition accuracy, energy consumption, obtrusiveness, and flexibility. Wang and Zhou [227] surveyed radio-based HAR methods, mainly including Zigbee-based, WiFi-based, and **Radio Frequency Identification– (RFID)** -based methods. Shoaib et al. [196] reviewed HAR systems that are implemented on smartphones and use only on-board phone sensors. Yousefi et al. [244] conducted a survey on HAR using WiFi **Channel State Information (CSI)**. Mukhopadhyay [141] reviewed relevant technologies and methods of human activity monitoring with wearable sensors. Bulling et al. [27] provided a tutorial on HAR with conventional machine learning methods based on wearable inertial sensors. Guo and Lai [74] conducted a survey on human action recognition with still images. While these existing survey works have reviewed many HAR methodologies and systems from different perspectives, they focused mainly on the taxonomy of HAR and reviewed the state-of-the-art HAR systems implemented with conventional machine learning methods.

In this article, we review systematically the techniques and methods related to HAR with deep learning. This study first presents a taxonomy of human activity, and then introduces commonly used sensors, preprocessing techniques, deep model building, and evaluation techniques. Although several recent works [151, 224, 250] have reviewed some works using deep models for activity recognition, they cover few deep models and do not involve preprocessing techniques nor the evaluation methods and metrics. There is still a need for a complete, comprehensive survey on HAR with recently developed deep learning methods, which is the main motivation of this work.

This survey is structured as follows: Section 2 introduces the human activity types and sensors that have been studied in the literature. Section 3 provides an overview of deep learning models for HAR. In Section 4, commonly-used data preprocessing techniques for deep learning are introduced. Section 5 presents core deep models and their variants, as well as their applications in HAR. Section 6 introduce evaluation methods, metrics, and public datasets, respectively. Finally, we conclude this article in Section 7 and give open research challenges.

2 BASICS OF HUMAN ACTIVITY RECOGNITION

2.1 Activity Category

While describing a specific human motion that has been done, it is common that one may find confused of choosing the appropriate term. Action, activity, and behavior are all frequently used candidates. Though they may seem to be similar, referring to the Oxford English Dictionary: *Action* means something that is done; *Activity* represents the state of being actively occupied, brisk,

Table 1. Categories of Activities Studied in the Literature

Category	Specific Activities
Locomotion	Walking, running, jogging, lying, standing, sitting, going upstairs/downstairs, and so on
Transport mode	Cycling, riding a bus, driving, traveling with a vehicle, and so on
Phone usage	Texting, making a call, using an app, browsing the web, checking the email, and so on
Entertainment	Playing soccer, playing basketball, attending a party, gaming, and so on
Health-related activity	Falls, respiration, rehabilitation activities, smoking, and so on
Daily activity	Sleeping, using computer, shopping, eating, attending a meeting, having a conversion, going to work, and so on
Gesture	Body gestures, arm gestures, hand gestures, head gestures, body languages, sign languages, and so on
Emotion	Angry, disgust, fear, happy, sad, surprise, neutral, and so on
Security	Presence, attacking, abnormal activities, and so on

This table is adapted from Reference [93].

or vigorous action; whereas *Behavior* depicts the manner of conducting oneself in the external relations of life. Each of them has its particular meaning and can derive different sets of jargon.

In philosophy and sociology, the three terms can be further defined. Max Weber described the connection between action and social action [30]. He believes action is motivated by an actor's feelings or experiences and is done on purpose. However, behavior can only be considered as a reflection of cues or impulses. Campbell [30], citing Weber's work, sees that action is an intentional activity, whereas behavior is an activity done without purposes or intentions. Action requires consciousness of the actor, which means the actor is intended to perform a certain activity, yet behavior requires a cue or stimulus to trigger. The relationship is thus clear. Action can be considered as a proactive activity, and behavior can be seen as a reactive activity. Beyond the two is the term activity, which is the super-set of both.

In addition, activity can be categorized into different levels, from low to high, according to its complexity [138]. For the sake of unambiguity, in this article, we do not differentiate activities of varying levels. Instead, we extend the taxonomy of activity in Reference [93], and categorize human activities into different types according to application domains. Table 1 shows the main categories of activities that are studied in the literature, including locomotion (e.g., walking, standing, running), transport mode (e.g., cycling, driving), phone usage (e.g., texting, making a call), entertainment (e.g., playing basketball, attending a party), health-related activities (e.g., falls, respiration), daily activities (e.g., sleeping, eating, going to work), gesture (e.g., hand gestures, arm gestures), and security (e.g., presence, attacking).

2.2 Sensors Used for HAR

There are a variety of sensors that can be used for HAR, which are mainly categorized as ambient sensors, wearable sensors, and other sensors.

2.2.1 Ambient Sensors. Ambient sensors require to be installed at fixed locations to recognize activities, which usually contain a server (e.g., WiFi access point) and a client (e.g., WiFi receiver in a smartphone). For simplicity, we call both servers and clients as sensors without distinguishing them. Ambient sensors that have been used for HAR mainly include **Global Navigation Satellite System (GNSS)**, Cellular, WiFi, Zigbee, FM (Frequency Modulation), and RFID.

- **GNSS**. The GNSS module built in smart devices can be used for positioning as well as activity detection. By using the location, moving speed, and number of available satellites provided by the GNSS, it is possible to recognize locomotion, transportation mode, and daily activities. For example, Liao et al. proposed a discriminative relational method for recognizing human activities solely using GPS data based on the **Relational Markov Networks (RMN)** framework. The proposed method performed efficient inference and learning with MCMC algorithms in extended RMNs [124]; Zhen et al. proposed a method based on supervised learning for inferring people's motion modes from GPS logs, such as the modes of walking and driving [258]; and Zheng et al. reported on an automatic inference inferring approach for transportation modes using supervised learning from raw GPS logs. The proposed method enabled to identification the modes of walk, driving, bus and bike, containing three parts, i.e. a change point-based segmentation method, an inference model using and a graph-based post-processing algorithm [257].
- **Cellular**. A cellular network is a communication network that is commonly used in mobile phones. Popular cellular technologies include GSM, CDMA, GPRS, UMTS, LTE, and so on. By measuring the **Received Signal Strength (RSS)** between the cellular receiver and transmitter, one can recognize different activities such as transport mode, daily activies. For instance, Sohn et al. investigated a system to recognize high-level properties of user mobility and daily step count based on coarse-gained GSM data from mobile phones [200]; Anderson and Muller developed a method to identify activities (e.g., walking, traveling in a motor car and remaining still) based on information readily available on a typical GSM cell phone, which can realize a context awareness level similar to that of an accelerometer [9].
- **WiFi**. WiFi is a local-area wireless communication technology that sends signals from a transmitter to a receiver. The available measurements of WiFi for HAR include RSS, CSI, and **Round Trip Time (RTT)**. Due to the ubiquity of WiFi infrastructures in urban and indoor environments, it has been widely used for positioning and activity recognition. For example, Sigg et al. investigated the use of WiFi RSS at a mobile phone for the recognition of situations, activities, and gestures [197]. Wang et al. proposed a device-free activity recognition system with deep learning approach based on CSI information [222]. In recent years, many works have used WiFi CSI for recognizing gestures [2, 248] and other types of activities [150, 223, 236, 242].
- **Zigbee**. Zigbee is a communication technology that is intended to be simpler and less expensive than Bluetooth, and WiFi. It uses IEEE 802.15.4 protocols to create personal area networks with small, low-power digital radios. Scholz et al. showed the general feasibility of activity recognition using Zigbee RSS on simple transceiver hardware [188]. Qi et al. developed RadioSense, a prototype system that exploits wireless communication patterns for activity recognition [167].
- **FM**. FM radio is a technology that conveys information by varying the frequency of a carrier wave. Shi et al. demonstrated that human activity recognition can be done with ambient FM-radio signals. They used fluctuations in the ambient signals of FM radio stations to distinguish empty room, opened door, and walking person [194, 195].
- **RFID**. RFID is a commonly-used technique to automatically identify and track tags by detecting the electromagnetic pulse from a nearby reader. It can be used for HAR, since the movement of human would change the single strength received by the reader. Li et al.

present a system for activity recognition from passive RFID data using a deep convolutional neural network [118]. Liu et al. proposed to use RF tag arrays and data-mining techniques for activity monitoring [127]. Wang et al. quantified the correlation between RF phase values and human activities by introducing TACT and modeling intrinsic characteristics of signal reflection in contact-free scenarios [233]. Fan et al. introduced an advanced RFID activity identification framework, DeepTag, which used a deep learning-based approach that combined a convolutional neural network and **Long Short-term Memory (LSTM)** network for activity identification in multipath-rich environments [56].

The advantages of using ambient sensors for HAR include non-intrusiveness toward users, and support multi-occupant activity detection. However, they have poor coverage and the achieved accuracy is influenced by many factors (e.g., people movement, obstacles).

2.2.2 Wearable Sensors. Wearable sensors are the sensors that are easy to carry and are usually built in smart devices. Common wearable sensors used for HAR are accelerometer, gyroscope, magnetometer, barometer, camera, acoustic sensor, light sensor, and biosensor.

- **Accelerometer**. An accelerometer is a tool used to measure the acceleration of a body in its own instantaneous rest frame. It has been integrated into most modern smart devices (e.g., smartphones, smart watches). Ravi et al. illustrated their efforts on users' activity recognition from a single triaxial accelerometer worn near the pelvic region [175]. Chen et al. introduced a deep learning-based system called METIER for activity recognition, which was evaluated on several accelerometer-based datasets [35]. Pei et al. also presented a deep learning-based system, called MARS, for recognizing different locomotion activities using several inertial measurement units that are placed on different parts of the body [160].
- **Gyroscope**. A gyroscope is a device that measures orientation and angular velocity, which is often used with an accelerometer. Pei et al. utilized the gyroscope accelerometer data to distinguish different locomotion activities [159]. Gu et al. developed a deep learning method for locomotion activity recognition by utilizing multiple sensors data built in most smart devices [70]. Zhou et al. proposed a convolutional neural network-based method for pedestrian activity recognition using a gyroscope and other sensors [260].
- **Magnetometer**. A magnetometer is a device used for measuring magnetic fields, which is also often used with a accelerometer. It has been widely used for activity recognition in many works [70, 159, 260].
- **Barometer**. A barometer is a device that measures the air pressure in a certain environment. Gu et al. used the barometer data for recognizing going upstairs, going downstairs, taking an elevator upward or downward [68, 69, 193]. Ye et al. proposed a barometer-based floor localization system, which uses the barometer of a smartphone to identify the floor of a mobile user in a multi-floor building [243].
- **Camera**. A camera is an optical instrument that captures images, from which a lot of activity types can be recognized. Nagarajan et al. introduced an environmental affordance model that learns directly from the egocentric video, primarily gaining a human-centered model of physical space (e.g., kitchen) that captures the primary spatial zones of interaction and the likely activities they support [145]. Tang et al. studied the RGB-D egocentric action recognition problem. The self-generated egocentric video is generated by wearable sensors, and the deep neural network method was used to explore the shared information and features of different modes [211]. Wang et al. proposed a novel symbiotic attention framework leveraging privileged information for egocentric video recognition [231].

- **Acoustic sensor**. An acoustic sensor is used to measure sound levels. Wang et al. proposed a contact-free acoustic gesture recognition system that adopts a frequency-hopping mechanism to mitigate frequency selective fading and avoid signal interference [232]. Ling et al. proposed an ultrasonic finger motion perception and recognition system based on **Channel Impulse Response (CIR)**. The system uses CIR measurements as gesture recognition features and utilizes a CNN model to classify the acquired images into different gestures [125]. Sun et al. proposed a system that supports fine-grained gesture-sensing on the back of mobile devices based on acoustic signals, which uses both the structure-borne and the air-borne acoustic signals to measure touch gestures [204].
- **Light sensor**. A light sensor is a photoelectric device that converts light energy into electricity, which has been integrated into many smart devices. Zhou et al. designed a system called IODetector to implement indoor/outdoor environment detection using the smartphone built-in light sensor [261]. Choudhury et al. proposed an automatic activity recognition system using light sensors [44].
- **Biosensor**. A biosensor is an analytical device used to evaluate and record the electrical activity of the human's muscles, heart, and so on. Popular biosensors are Electromyography and Electrocardiography, which measure the electrical activity of muscles, and that of the heart, respectively. Song et al. proposed an end-to-end spatial and temporal attention model for human action recognition from skeleton data [201]. Zheng et al. used machine learning methods to study emotion recognition over time in stable patterns of **Electroencephalogram (EEG)** patterns [255].

Compared to ambient sensors, wearable sensors do not suffer the coverage problem, and can achieve a higher accuracy as the user usually carries the sensors. However, wearable sensors can only detect the activity of the person who carries them, and do not support multiple person detection. Apart from ambient sensors and mobile sensors, there are some other sensors, such as event camera, which can also be used for HAR.

2.2.3 Other Sensors.

- **Event camera**. An event camera is a relatively new sensor that captures the brightness change of the scene. Different from conventional cameras, which output synchronous frames, event cameras output asynchronous events. Amir et al. proposed an event-based gesture recognition system using a TrueNorth neurosynaptic processor [8]. Miao et al. introduced several event-based datasets for pedestrian detection, action recognition, and fall detection [134].

Table 2 summarizes the sensors that have been used for HAR in the literature.

3 OVERVIEW OF DEEP LEARNING FOR HAR

In this section, we provide an overview of using deep learning methods for HAR. As shown in Figure 2, there are four components for deep learning–based HAR systems. First, data are collected from a variety of sensors, and these data can be images, WiFi CSI, accelerations, gyroscope readings, barometer readings, sound, biosensor readings, and so on. Second, the input data are preprocessed using certain techniques such as scaling, **Principal Component Analysis (PCA)** whitening, **Zero-phase Component Analysis (ZCA)** whitening, or denoising. Third, the preprocessed data are fed to model building component where different deep models (e.g., RBM, autoencoder, RNN) can be chosen to learn useful features. This is followed by a classifier at the top layer (e.g., softmax classifier, SVM). Once we build a model, we can train it with the input data. During

Table 2. Sensors used for HAR in the Literature

Sensor	Reference	Activity Category
GNSS	[124, 257, 259]	Locomotion, transportation mode, daily activity
Cellular	[9, 200]	Locomotion, transportation mode, daily activity
WiFi	[2, 150, 197, 222, 236, 242, 248]	Locomotion, transportation mode, daily activity, gestures
Zigbee	[167, 188]	Locomotion activity
FM	[194, 195]	Locomotion, daily activity
RFID	[56, 118, 127, 233]	Locomotion, daily activity, security
Accelerometer	[35, 160, 175]	Locomotion, phone usage, entertainment, gesture, transportation mode, daily activity
Gyroscope	[70, 159, 260]	Locomotion activity
Magnetometer	[70, 159, 260]	Locomotion activity
Barometer	[68–70, 193, 243]	Locomotion activity
Camera	[145, 145, 201, 211, 211, 231]	Locomotion, entertainment, daily activity, health-related activity, gesture, security
Acoustic sensor	[125, 198, 204, 232]	Locomotion, transportation mode, daily activity
Light sensor	[44, 261]	Locomotion activity, gesture
Biosensors	[201, 255]	Locomotion, health-related activity, emotion, daily activity
Event camera	[8, 134]	Gesture, locomotion activity

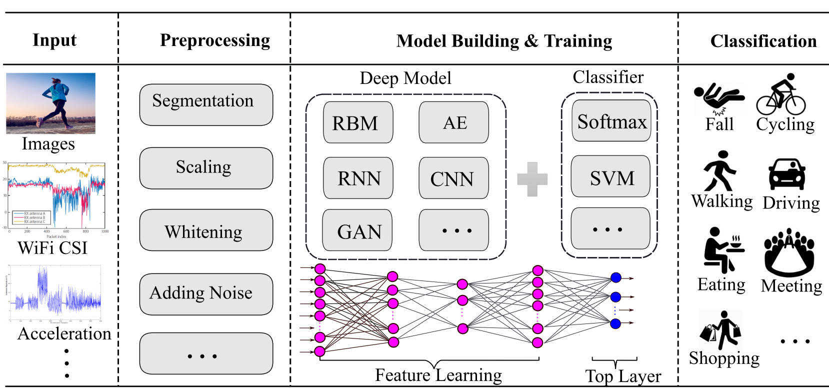


Fig. 2. Overview of deep learning for HAR.

the training process, the network parameters such as weights are optimized. Finally, we can use the trained model to predict the activity of upcoming data.

In the following, we will elaborate the techniques related to preprocessing, model building, and evaluation.

4 PREPROCESSING TECHNIQUES

Before we feed data to a deep model, certain preprocessing techniques are required to be used to obtain satisfactory performance. The main preprocessing techniques include segmentation, scaling, one-hot encoding, dealing with missing data, transformation, adding noise, and denoising. We will introduce each of them in the following.

4.1 Segmentation

Depending on the used sensor data, segmentation might be necessary to be conducted on the data before feeding it to a deep model. For image-based HAR, it is feasible to recognize activities such as gestures and walking from a single image. However, we cannot recognize activities from a single accelerometer reading, barometer reading, WiFi CSI, and so on. This is because a single data sample (except an image) cannot capture the characteristics of an activity. Therefore, we need to segment these data into sequences using a fixed time window (e.g., 2 s), on which we can build models for HAR. Note that when combining data from different sensors for conducting HAR, we need to align them to the same time window as the sampling frequency for different sensors might be different. It is also usually required by some deep models (e.g., autoencoder) to stabilize the number of input samples within a time window through interpolation (e.g., spline interpolation), since the sampling rate for the same sensor may not be stable [70]. In recent years, some time-aware models have been proposed to deal with irregular time intervals in data [18].

4.2 Scaling

Raw data are usually not useful enough for machine learning methods unless the raw attributes have a meaning in the original domain [60]. To enable deep models to achieve desirable performance, we usually require to rescale the raw data to a certain range, since deep models are usually preferable to work on inputs of small values (e.g., between 0 and 1). If the input values of a model are too large, then the model tends to learn large weight values, which increases the computational cost and may lead to overflow on digital computers [64]. Two common scaling techniques are normalization and standardization.

Normalization is a technique that rescales the data from the original range to the range between 0 and 1. Let $\boldsymbol{x}$ indicate an input vector of sensor readings, which usually corresponds to one column of the input matrix. The normalization process is mathematically expressed as follows:

$$\boldsymbol{x}' = \frac{\boldsymbol{x} - min(\boldsymbol{x})}{max(\boldsymbol{x}) - min(\boldsymbol{x})}, \tag{1}$$

where *min* and *max* are the functions to calculate the minimum and maximum values of the input vector, respectively. The resulting vector $\boldsymbol{x}'$ ranges from 0 and 1. However, normalization may not work well in some cases where the maximum or minimum value of the vector is not available or there are extreme outliers [60]. For simplicity, we also use $\boldsymbol{x}'$ to represent the resulting vector after a manipulation in the following.

Standardization is another popular scaling technique, which is less influenced by the presence of outliers. It is mathematically described as

$$\boldsymbol{x}' = \frac{\boldsymbol{x} - \mu_x}{\sigma_x}, \tag{2}$$

where μ_x and σ_x are the mean and standard deviation of the values of the input vector $\boldsymbol{x}$. After the standardization manipulation, the resulting data have a mean of 0 and a standard deviation of 1. If the mean and standard deviation of the corresponding probability distribution are not available, then the sample mean and standard deviation will be used instead [60].

4.3 Label Encoding

Activity labels, such as walking and shopping, are usually categorical, but deep models cannot work on the categorical labels efficiently, and require all the input data to be numeric. To do that, we can simply encode each label as an integer. However, integer encoding may not perform well as the model may try to learn a natural ordering relationship in categories. To avoid this, a more common way is to encode the label with one-hot encoding [142]. In the one-hot encoding, an identity matrix with the size equal to the number of activity categories is used. Each row represents an activity, and has and only has one element with the value of 1 that indicates the activity.

4.4 Dealing with Missing Data

Missing data can lead to efficiency loss, biased results, and complexity increase [60]. There are different ways to deal with missing data. A simple way is to discard the samples with missing values in their attributes or even to delete the entire attribute if any of samples has a missing value. However, this may significantly reduce the amount of training data if the number of samples with missing values is large, which will decrease the performance of the classifier trained on it. An alternative way is to replace the missing values with the mean/median of the non-missing values in the same attribute. While it is easy and fast, it does not consider the correlations between attributes. A more sophisticated way is to do imputation that replaces missing values with the values estimated by imputation methods. The commonly used imputation methods include K-nearest neighbors [55], maximum likelihood imputation [60], singular value decomposition [215], and fuzzy k-means clustering [116]. More details of imputation methods can be found in the book [60].

4.5 Transformation

It is often advantageous to conduct certain transformations on the input data before using them to train a deep model. The transformations are usually used to reduce the correlations in the input data. Popular transformations include PCA whitening, ZCA whitening, and spectrogram analysis.

4.5.1 PCA Whitening. Whitening (or sphering) is a linear transformation that converts the input vector into another vector with the unit diagonal white covariance, which can be viewed as a generalization of standardization [101]. PCA whitening is one of commonly-used preprocessing methods reduce the redundant information in the input data so that the deep models can learn features more efficiently [122, 146]. Since PCA whitening is based on the PCA method, it enables us to obtain whitened data with lower dimensionality than the original input by simply keeping top k components [147, 220].

4.5.2 ZCA Whitening. ZCA whitening is another popular preprocessing method to reduce the correlations in the input data. ZCA whitening is related to PCA whitening [101], and the ZCA whitening matrix is obtained by multiplying an orthogonal matrix with the PCA whitening matrix. Different from PCA whitening, ZCA whitening does not reduce the dimensionality of inputs. ZCA whitening has been widely used in image processing [39, 105, 122, 146], and activity recognition [70]. Apart from PCA whitening and ZCA whitening, there are other whitening methods such as Cholesky whitening, ZCA-cor whitening, and PCA-cor whitening [101].

4.5.3 Spectrogram Analysis. As many types of sensor data for HAR (e.g., accelerometer readings) are usually time series, spectrogram analysis might be helpful in capturing variations in the input data. A spectrogram is a representation of the frequency spectrum of the input signal as it varies with time, which can be generated by the Fourier transform [25] or a wavelet transform [191]. In the spectrogram analysis using short-time Fourier transform, the spectrogram can be considered as the squared amplitude of a time-frequency transformation of the signal (e.g., a time-frequency energy density function) [191]. Spectrogram analysis has been used extensively in the fields of speech processing [254], image processing [228], music [81], and others. In Reference [6], the spectrogram of an accelerometer signal has been used for deep learning-based activity recognition and witnessed better classification accuracy and lower computational complexity. In Reference [70], apart from spectrogram analysis, more frequency-related transformations have been conducted for activity recognition, including fast Fourier transform, power spectral density, discrete cosine transform, and cepstrum analysis. It concludes that using spectrogram can result in better accuracy on accelerometer data-based classification, but other transformations do not lead to a better accuracy than the original data.

4.6 Adding Noise

Adding noise to input data is a commonly used way to help deep learning models learn good representations that are rather robust under corruptions of the input. There are three popular types of noise used for preprocessing sensor data [219]: (1) additive Gaussian noise where a Gaussian noise term is added to the input; (2) masking noise where a fraction of the input's elements is set to zero; (3) salt-and-pepper noise where a fraction of the input's elements is assigned to their maximum or minimum value. Additive Gaussian noise is a common noise model for real-valued inputs, while the salt-and-pepper noise is considered when input domains are binary or near binary. Masking noise is often regarded as turning off components, which numerically means forcing components to zero so that these components are ignored in the computations of downstream neurons. Note that these noise can be added not only to the inputs, but also to activations, weights, gradients, and outputs. These operations are usually helpful for deep models to achieve good accuracy on the HAR task.

4.7 Denoising

Denoising is to remove or reduce the noise in the input data so that the model can achieve better recognition accuracy. A commonly used denoising method is to apply filters (e.g., low-pass filters, median filters, Kalman filters) to the input data. For example, low-pass filters are often applied to the acceleration signal to remove the gravitational and body motion components [79]. Median filters are good at removing the salt and pepper noise of images for vision-based activity recognition [11]. In Reference [237], the authors used a series of denoising methods (low-pass filtering, PCA, and median filtering) to deal with CSI-based activity recognition. In Reference [229], the authors compared the performance of different filters for accelerometer-based activity recognition, and concluded that the Kalman filter had the largest signal-to-noise ratio, followed by median filter, and low-pass filter. Apart from using filters, some studies reduce noise by applying a threshold to the data. In Reference [61], the authors first extracted histograms from accelerometer data for each activity, and then reduced noise by setting a threshold to the corresponding histograms.

4.8 Summary

Preprocessing is a crucial step of using deep learning for HAR, which may lead to low classification performance if it is not done properly. As for which preprocessing method should be used, it depends on the sensor data used, the selected deep model, as well as the activities of interest.

One important thing regarding preprocessing is that any preprocessing methods must only be conducted on the training data and then applied to the validation or test data [117]. For example, it would be a mistake to first compute the mean of input data across the entire dataset and then split the data into training and test data.

5 DEEP LEARNING MODELS

Recent years have witnessed the rapid development and advances of deep learning, and many deep models have been developed in the literature. In this section, we are interested in popular deep models that are constructed on some common building blocks including RBM, AE, CNN, RNN, and GAN, and some of their variants. The deep models based on these building blocks have been used or have the potential to be used for HAR.

5.1 RBM-based Models

RBM-based deep models are one of early successful deep models used for HAR, which are obtained by stacking multiple RBMs. The main intuition of applying RBM-based models for HAR is to reduce

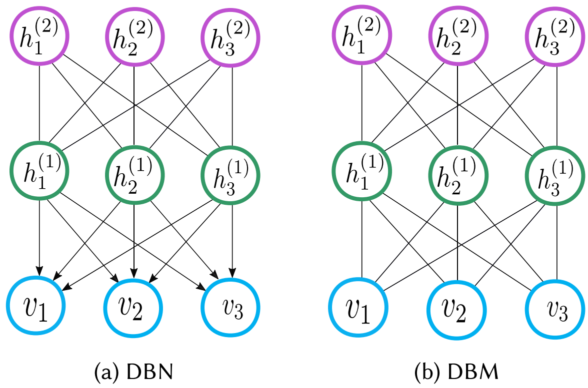


Fig. 3. (a) The structure of a DBN where there are multiple hidden layers and have both directed and undirected connections. Apart from the connections between visible units and hidden units, it has also connections between hidden units in different layers. (b) The structure of a DBM where there are undirected connections not only between visible units and hidden units but also among hidden units in different layers.

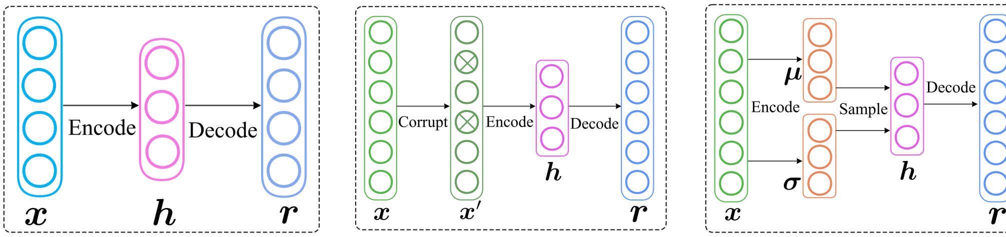
the dimensionality of sensor data and to extract useful features in an unsupervised way. Popular deep models based on RBMs include **Deep Belief Networks (DBNs)** [19, 84], **Deep Boltzmann Machines (DBMs)** [185], **Convolutional Boltzmann Machines (CBMs)** [115], and so on. We briefly introduce the DBN and DBM in the following.

A DBN has multiple hidden layers, as shown in Figure 3(a). The visible units in the DBN may be binary or real, while the hidden units are usually binary. Typically, the units in one layer are fully connected to the units in the neighboring layers except in a sparse DBN. The connections between the top two layers are undirected, while those between all the other layers are directed [64]. A DBN can be constructed by stacking multiple RBMs [84, 140].

DBNs are one of the first successful deep models for HAR. For instance, Zhang and Wu [252] presented a DBN-based method for activity detection based on voice signals. Fang and Hu [57] utilized a DBN of four hidden layers to recognize daily activities in a smart home. Uddin et al. [217] introduced a DBN method that consists of three hidden layers for recognizing facial expression. Zheng et al. [256] developed an EEG-based emotion method using a DBN. DBNs have a distinguishable property from other directed generative models that allow to infer the states of hidden units in a single forward pass [84]. The resulting weights can be used to initialize all the feature detecting layers of a classification network. However, DBNs are rarely used nowadays due to the problems associated with both directed models and undirected models such as intractable inference to marginalize out the hidden units, and intractable partition function of the top two layers [64].

Another popular deep model based on RBMs is DBM [185], which is also a generative model. Similar to DBNs, DBMs also consist of multiple RBM but all the connections in the DBMs are undirected, as shown in Figure 3(b). DBMs can be represented as a bipartite graph, which enables the conditional distribution over one DBM layer to be factorial. Compared to DBNs, DBMs are simpler, but allow richer approximations of the posterior [64]. In DBNs, there exist a series of variational bounds on the log probability of the training data [85], which cannot be explicitly optimized. By contrast, it is feasible to actually optimize the variational bounds in DBMs, since all the hidden units in one DBM layer are conditionally independent given the other layers.

So far, some works based on DBMs have been done for HAR. Bhattacharya and Lane [24] used a three-layer model consisting of RBMs to recognize gestures, transportation mode, and



(a) AE (b) DAE (c) VAE

Fig. 4. (a) The structure of an AE. It is trained to minimize the error to reconstruct the original input from the features (representations). (b) The structure of a DAE. It takes as input the corrupted version of input and is trained to reconstruct the original input. (c) The structure of a VAE with Gaussian probability distributions in its latent space. It is trained to learn the probability distribution of data, from which features are sampled and used to reconstruct its original inputs.

indoor/outdoor activities. Plötz et al. [162] presented a DBM-based method to learn features from data automatically for activity recognition. Lane et al. [109] proposed a deep model that is composed of three layers of RBMs for different audio sensing tasks (e.g., ambient scene analysis, stress detection, emotion recognition, and speaker identification). Radu et al. [172] presented a multimodal DBM learning method for HAR on mobile devices. Overall, RBM-based deep models are rarely utilized for HAR these days due to the difficulty to train them.

5.2 Autoencoders

AEs are another type of generative models, which are similar to RBM-based models. While both of them can learn useful representations of original data [21, 216], the AEs use deterministic units instead of stochastic units that are used in RBMs. An AE has two processes: encoding and decoding. In the encoding process, original inputs are transformed into features (representation of data). In the decoding process, the learned features are reconstructured to approximate the original inputs. By minimizing the reconstruction error between the input data and its reconstruction, the learning of an AE is trained. Figure 4(a) shows the encoding and decoding processes.

Like PCA, autoencoders were originally proposed for dimensionality reduction. An autoencoder with the input dimensionality larger than the feature dimensionality is called an undercomplete autoencoder [21]. The undercomplete autoencoder with linear decoder and the mean squared loss function learns the same subspace as PCA, but the autoencoder that has nonlinear encoder and decoder functions can learn a more powerful generalization than PCA. While undercomplete autoencoders are able to learn the most salient features of data, they may simply fail to learn useful information if the encoder and decoder have too much capacity [64]. Overcomplete autoencoders, where the feature dimensionality is greater than the input, also suffers a similar problem. To solve the problem, regularized autoencoders have been proposed in recent years, including **Sparse Autoencoder (SAE)** [148], **Denoising Autoencoder (DAE)** [219], **Variational Autoencoder (VAE)** [83, 103, 238], and so on [216].

Due to the excellent performance and representation ability, AE and its variants have been widely used for HAR. Gu et al. [70] designed a HAR method based on an stacked DAE based on data from four types of sensors built in a smartphone, including accelerometer, gyroscope, magnetometer, and barometer. Almaslukh et al. [5] developed a HAR method based on a SAE for recognizing locomotion activities. Wang [226] presented a HAR method based on a continuous AE, which uses data from accelerometer, gyroscope, and magnetometer. Li et al. [121] investigated the

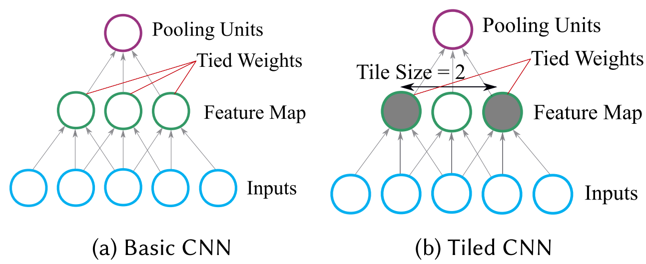


Fig. 5. (a) A basic CNN with local receptive fields and units in each feature map have the tied weights. (b) A tiled CNN with partially untied local receptive fields and units with the same texture in each feature map have tied weights. This figure is adapted from Reference [111].

performance of SAE, DAE, and PCA in feature learning for activity recognition. Hasan and Roy-Chowdhury [78] introduced a framework based on stacked SAEs and active learning to recognize activities from streaming videos. Wang et al. [225] utilized a SAE to simultaneously recognize location, activity, and gesture from wireless signals. Wang et al. [230] presented a stacked AE for egocentric activity recognition from videos. Zhang et al. [251] utilized an AE to learn features from electroencephalographies to recognize brain activities. In HAR, AE and its variants are mainly used to reduce data dimensionality and learn features from data. Apart from achieving excellent performance, they can also make use of unlabeled data. They are often used for processing data from inertial sensors, WiFi, acoustic sensor, and so on. However, they are rarely used for dealing with images where CNNs methods perform the best.

5.3 Convolutional Neural Networks

Different from AEs and RBM-based methods, CNNs are discriminative models that use convolution operation to replace general matrix multiplication in at least one of their layers [80, 106, 113, 208]. CNNs include two operations: convolution and pooling. Convolution involves three key ideas: sparse connectivity, parameter sharing, and equivariant representations. Pooling uses a statistic of the inputs as the output. There are different pooling functions such as max pooling [106], average pooling [113], L^2-norm pooling [64], and tree pooling [114].

Basic CNNs usually have four types of layers: convolution layer, pooling layer, detector layer (e.g., ReLU layer), and fully connected layer (a layer in general neural networks). These layers can be stacked to form a deep CNN. Due to the excellent performance of CNNs in different domains especially image classification, many variants of CNNs have been proposed [72]. One of the popular variants is called tiled CNN [111]. While weight sharing mechanism in convolution can significantly reduce the number of parameters, it restricts the model from learning other invariant features (e.g., rotational invariant features). A solution to address this problem is the tiled CNN, which can learn diverse feature maps by constraining weights functions k steps away from each other to be equal. The parameter k here is called tile size, and the tiled CNN corresponds to the basic CNN when the tile size equals to 1. Compared to basic CNNs, tiled CNNs can not only reduce the number of parameters to be trained, but also allow to learn other invariances, as shown in Figure 5(a) and (b). It has been demonstrated that tiled CNNs outperform basic CNNs in References [111, 234]. Apart from tiled CNNs, there are many other variants of basic CNN. In Reference [72], the authors introduced many variants of CNN that are improved from different aspects including layer design, activation function, loss function, regularization, computation, and optimization.

As one of earliest successful deep learning methods, CNNs have also been widely used for HAR. Ronao and Cho [182] used a CNN to recognize six types of locomotion activities and demonstrated

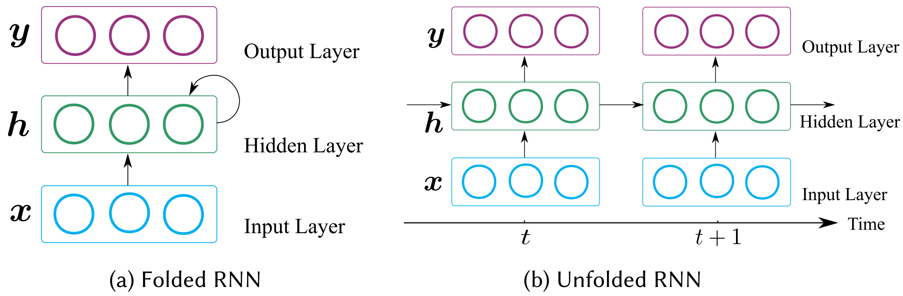


Fig. 6. Examples of folded RNN and unfolded RNN. (a) A simple folded RNN where there are loops in the hidden layers. (b) A simple unfolded RNN. The hidden units of the network at different times have connections.

that the CNN outperforms some conventional methods such as MLP, Naive Bayes, and SVM. Hughes and Correll [90] applied a distributed CNN to recognize some mid-level activities (e.g., opening a door, opening a drawer), and analyzed the effect of sensor location (e.g., body, arms, and legs) for recognizing these activities. Yang et al. [241] designed a CNN for recognizing hand gestures and daily activities. Ha and Choi [75] presented a CNN that employs both partial weight sharing and full weight sharing for HAR based on multimodal data (specifically, data from multiple accelerometer and gyroscope sensors). Zeng et al. [246] introduced a CNN for HAR using mobile sensor data. Zhou et al. [260] proposed a CNN-based locomotion activity recognition method using data from accelerometer, magnetometer, gyroscope, and barometer.

Overall, CNNs are suitable for dealing with data with a known, grid-like topology (e.g., images, time-series data that can be considered as one-dimensional (1D) grid sampling at regular time intervals). However, CNNs may not work well on processing sequential data.

5.4 Recurrent Neural Networks

RNNs are a family of neural networks that have recurrent connections [64, 183, 186]. While CNNs are specialized for processing matrix-like data (e.g., images, video streams), RNNs are favorable for dealing with sequential data (e.g., speech, accelerometer readings). Figure 6 shows the structure of a simple folded RNN and unfolded RNN.

The recurrent connections in RNNs enable them to outperform general neural networks in dealing with sequential problems, since they can learn the sequential dependencies. However, the memory produced from the recurrent connections is often limited by the algorithms used to train the RNN due to vanishing or exploding gradients issues. A popular way to reduce the effects of vanishing and exploding gradients is to use LSTM RNN [66]. Different from conventional RNNs, the LSTM RNN replaces hidden units with memory cells whose inputs and outputs are controlled by gates to store or forget information [186]. The LSTM was originally proposed in Reference [87] and modified in Reference [66] that has become popular later.

While the standard LSTM has demonstrated promising performance in a variety of tasks, it may fail to understand the input structures that are more complicated than a sequence. To address this challenge, a tree-structured LSTM network is proposed in Reference [263], which is called S-LSTM. The S-LSTM network consists of S-LSTM memory blocks, including an input gate, two forget gates, a cell gate, and an output gate. While the S-LSTM network can achieve better performance in complicated sequential modeling problems than the standard LSTM network, it has higher computational complexity. Another notable variant of the basic RNN is **Gated Recurrent Units (GRUs)** [40, 41], which can adaptively capture sequential dependencies. While it has a gated structure, which is similar to the LSTM, it is more computationally efficient than the LSTM. The GRU shares some common characteristics with the LSTM [45]. The most salient one is that both have the additive component in the update process from time t to $t + 1$, which distinguishes them from traditional recurrent units. This characteristic allows them to remember existing features and to bypass multiple temporal steps. However, they also have some differences. For example, the LSTM unit can control the exposure degree of the memory content by the output gate, while the GRU simply exposes the full content without any control. More specific similarities and differences between the LSTM and the GRU can be found in Reference [45]. In addition to the LSTM, GRU, and their variants, there are many other variants of RNN. Some of them are recurrent CNN [123], structurally constrained RNN [136], ubitary RNN [13], and gated orthogonal recurrent unit [96]. The advantages and disadvantages of these architectures are summarized in Reference [186].

RNNs have been widely used for HAR as activity recognition can be considered as a sequential problem. Usually, RNNs are considered as one kind of discriminative models though they can be used as generative models [206]. In the context of HAR, the discriminative RNNs are used. Its training is done in a supervised way, which minimizes the cost function of the network output and the corresponding label. Murad and Pyun [143] utilized a **deep RNN (DRNN)** consisting of LSTMs to recognize the activities from several open public datasets. They demonstrated that the unidirectional DRNN outperforms the bidirectional DRNN and the cascaded DRNN. Similarly, Inoue et al. [94] also used a DRNN composed of LSTMs to recognize human activities from acceleration signals. Guan and Plötz [73] developed a HAR method that ensembles multiple LSTMs, and showed that the ensembles of LSTM networks outperform individual LSTM networks. Qi et al. [166] proposed a structural-RNN method for recognizing group activities from videos, which are based on spatio-temporal attention and semantic graph. Edel and Köppe [54] developed a binarized BLSTM RNN for recognizing daily activities and locomotion activities. By replacing the arithmetic operations with bit-wise operations, the binarized BLSTM can significantly reduce memory size and accesses and is hence more computationally efficient than the standard LSTM. Overall, RNNs are perfect for processing sequential data due to its memory mechanism, but they consume larger energy to train.

5.5 Generative Adversarial Networks

GANs are a relatively new type of generative models [65], which have gained great popularity in image processing [126], classification [152], image generation [23], and so on. A GAN consists of a generator and a discriminator. The main task of the generator is to learn the data distribution so that it can generate samples to deceive the discriminator. By contrast, the discriminator examines samples to recognize whether they are from real data (training data) or fake data (produced by the generator). The training of the GAN is to let the generator and the discriminator to play a two-player zero-sum game until reaching a Nash equilibrium where the discriminator cannot tell whether the input sample is from the generated data [157]. Figure 7 shows the structure of a basic GAN that includes a generator and a discriminator.

GAN has achieved great success in different domains, but the original GAN also suffers from several problems such as gradient vanishing, poor diversity, and unstable training. Therefore, many variants of the original GAN [65] have been proposed. One of representative variant of the basic GAN is **Conditional GAN (CGAN)** [137], which controls the data generation process with auxiliary information (denoted by $\boldsymbol{c}$). Another representative variant of GAN is **Least Squares GAN (LSGAN)** [132], which uses the least squares loss function for the discriminator rather than the sigmoid cross entropy loss function used in the basic GAN. Compared to the basic GAN, the LSGAN can generate higher quality data and is more stable in the learning process. Apart from the CGAN and the LSGAN, there are other notable variants of GAN such as DCGAN [169], CycleGAN [262],

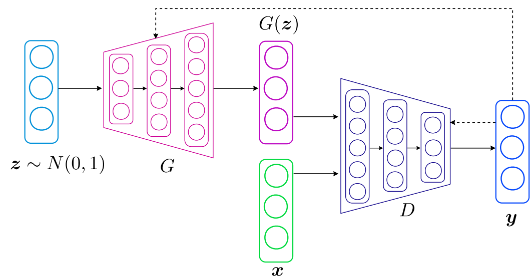


Fig. 7. The structure of a basic GAN, which contains a generator G and a discriminator D. The GAN is trained to update the parameters of G and D. The output $\boldsymbol{y}$ of the discriminator is a binary vector, indicating whether the input sample is real or fake.

InfoGAN [37], ProGAN [97], WGAN [12], SAGAN [247], EBGAN [253], BEGAN [23], and StyleGAN [98]. The key contributions, advantages, and disadvantages of GANs have been surveyed in References [47, 235].

To date, few works have used GANs and their variants for HAR, but they have great potential to be widely used in HAR as collecting labeled data in HAR is challenging and costly. The effort in collecting labeled data could be significantly reduced by making use of GANs. Li et al. [119] employed a CGAN to generate activity masks from video frames, which are then fed to a VGG-LSTM network [199] for activity recognition. Gammulle et al. [59] proposed a multi-level GAN architecture combining with LSTMs to recognize group activities from video sequences. Ahsan et al. [4] developed a HAR framework, which first trains a DCGAN on a large unlabeled video activity dataset and then fine-tunes the pretrained discriminator from the GAN model on a labeled dataset for activity recognition. Overall, GANs are suitable for the scenarios where labeled data are few. Existing GAN-based works for HAR use mostly videos or images. More studies are required to investigate the feasibility and performance of using GANs for HAR on other types of data (e.g., WiFi CSI, accelerometer readings).

5.6 Summary

Both generative deep models and discriminative deep models can be used for HAR. To date, there are many deep models available, it is natural to ask which model should one choose. While there is no consensus on this point, it might be useful to try different models that fit the problem domain. Discriminative deep models (CNNs, RNNs, and their variants) can be directly trained for HAR. CNNs are originally used for processing image and image-like data, and therefore they have excellent performance on HAR with images. Some recent works have also used CNNs on other sensor data (e.g., accelerometer data). RNNs are popular in dealing with sequential data, especially time-series data, and have been extensively used for HAR based on inertial sensor data.

Generative deep models (RBMs, AEs, and GANs) require to add a discriminative layer (e.g., softmax layer) on the top for HAR. They are often used to reduce the dimensionality of data or extract useful features in an unsupervised way. AEs are somehow similar to RBMs, but AEs are more straightforward and easy-to-understand than RBMs. Also, the optimization of AEs is much easier than that of RBMs. This makes AEs more popular than RBMs these days. GANs, which are another type of generative model, are often used to generate new images or other data. Although there are few works on using GANs for HAR, it is promising to apply GANs to reduce the number of labeled examples required to train a deep model.

Generative deep models and discriminative deep models can also be combined together to enjoy the benefits of both models. Such models are also called hybrid models. In Reference [192], an AE is combined with a CNN to classify human activities based on radar data. In Reference [264], an AE is fused with a CNN and a LSTM to achieve device-free HAR from WiFi CSI data.

Table 3 summarizes popular state-of-the-art HAR methods/systems with deep learning, including the model used, activity of interest, dataset, as well as their performance. From the table, one can see that varying deep models have been used for HAR. RBM and its variants (e.g., DBN, DBM) are widely used to recognize different activities in References [24, 57, 162, 172, 217, 256] and are gradually losing its popularity due to the difficulty in parameter optimization. AEs are also popular in HAR [5, 70, 78, 230, 251] and many variants have been developed. CNNs are becoming prevalent in HAR [75, 90, 182, 246, 260] nowadays though they are originally proposed for dealing with images. RNNs are more often used for modeling sequential data [73, 94, 143], which can capture the temporal dependencies of time-series data. GANs are less used in HAR so far, but they have great potential to be spread. Several works have integrated GANs with discriminative models (e.g., LSTM) [59, 119].

The most common activity type for which most deep models are trained is locomotion activity, followed by daily activity, gesture, and transport model. The may be because locomotion activity can be recognized by using both ambient sensors and mobile sensors. The ubiquity of sensor-rich mobile devices has made it easy to sense a variety of activity (not only locomotion activity, but also daily activity, gesture, and so on). Several studies have also investigated the use of deep models for recognizing emotion and health-related activity [109, 217, 256], which deserve more attention and efforts in the future. There is few work on investigating the applicability of deep models for recognizing activities such as phone usage, security, entertainment.

The commonly used datasets include Skoda [245], Opportunity [180], UCI-HAR [10], Ambient Kitchen [161], Daily Routines [91], PAMAP2 [177], USC-HAD [249], and USC-HAD [249]. More datasets that contain multimodal signals for more types of activity would be helpful in promoting the development and advances of HAR with deep learning.

Note that although deep learning networks have been widely used for HAR, there are scenarios that deep networks may not be the good choice. First, deep networks often require a large amount of training data to avoid over-fitting, which may be cost-prohibitive for some modalities (e.g., barometer data, and Zigbee data). Second, deep networks generally need to be run on high-performance servers or platforms, and therefore they are not suitable for resource-limited HAR (e.g., smartphone-based HAR tasks). Third, some types of data for HAR cannot be directly used by deep networks. For instance, the output of event cameras are discrete asynchronous events, and cannot be directly processed with state-of-the-art deep networks that usually run on fixed-size (e.g., grid-like) data.

6 EVALUATION

6.1 Evaluation Methods

There are mainly three methods used to evaluate the classification performance of a deep model, namely holdout, cross validation, and bootstrap [102].

6.1.1 Holdout. The holdout is a simple evaluation method, which randomly divides the dataset into training data, test data, and validation data (required by some modeling algorithms). The training data are used to fit the model. The validation data are utilized to assess the performance of the fit model to search for the optimal values of parameters. After that, the trained model is assessed on the test data to evaluate the generalization error of the model. The holdout method is

Table 3. State-of-the-art HAR Method/System with Deep Learning

Reference	Year	Deep Model	Activity	Dataset	Performance Metric	Performance
[24]	2016	DBM	Gesture, transport mode, indoor/outdoor detection	Private data collected by smart watch	Accuracy	About 73%, 93%, 94%, for respective activity
[162]	2011	DBN	Daily activity, factory maintenance	Ambient Kitchen [161], Daily Routines [91], Skoda [245], Opportunity [180]	Accuracy	88.7%, 86.8%, 75.8%, 74%, for respective dataset
[109]	2015	DBM	Daily activity, emotion	EmotionSense [168], StreeSense [129], SpeakerSense [128], JigSaw [130]	Accuracy	About 83%, 82%, 81%, 58% for respective dataset
[57]	2014	DBM	Daily activity	Private data collected by ambient sensors	Accuracy	About 87%
[217]	2017	DBN	Emotion	Private data collected by camera	Accuracy	96.67%
[256]	2014	DBN	Emotion	Private data collected by electroencephalography (EEG)	Accuracy	87.6%
[70]	2018	SDAE	Locomotion activity	Private data collected by smartphone	Precision, Recall F1 score	94% (F1 score)
[5]	2017	SAE	Locomotion activity	UCI HAR [10]	Accuracy	96.4%
[226]	2016	Continuous AE	Locomotion, transport mode	Altun et.al [7]	Accuracy	99.3%
[78]	2015	SAE	Locomotion, transport mode, daily activity	KTH [189], UCF50 [176], VIRAT [154], TRECVID [155]	Accuracy	98%, 53.8%, 62.6%, 6.7% for respective dataset
[225]	2017	SAE	Locomotion, gesture	Private data collected by WiFi nodes	Accuracy	85%
[230]	2018	AE	Locomotion, daily activity	UEC Park [104], LongEgo [163]	Accuracy, Recall	77.6% (Accuracy), 93% (Recall)
[182]	2016	CNN	Locomotion activity	Private data collected by smartphone	Accuracy	94.8%
[173]	2017	CNN	Locomotion activity, daily activity indoor/outdoor detection	STISEN [203], GAIT [149], Sleep-Stage [63], Indoor/Outdoor Detection [171]	F1 score	81.6%, 89.5%, 66.4%, 82.3% for respective dataset
[170]	2019	CNN	Locomotion activity	Private data collected by smartphone	Accuracy	92.38%
[241]	2015	CNN	Daily activity, gesture	Opportunity [180], Gesture [27]	F1 score, Accuracy	54.7% (F1), 89.6% (F1) for respective dataset
[260]	2019	CNN	Locomotion activity	Private data collected by smartphone	F1 score, Precision, Recall	97% (F1)
[221]	2019	CNN	Gesture	ARIL	F1 score, Precision, Recall	88% (F1)
[160]	2021	AE+CNN	Locomotion activity	Self-synthesized dataset	Accuracy	95.81%
[143]	2017	LSTM	Daily activity, locomotion, health-related activity, factory maintenance	UCI-HAR [10], USC-HAD [249], Opportunity [180], Daphnet FOG [14], Skoda [245]	Precision, Recall, Accuracy, F1 score	96.7%, 97.8%, 92%, 93%, 92.6% (Accuracy) for respective dataset
[94]	2018	LSTM	Locomotion activity	HASC [99]	Accuracy	95.4%
[244]	2017	LSTM	Locomotion activity	WiFi CSI	Precision	About 90.5%
[73]	2017	LSTM ensembles	Daily activity, factory maintenance	Opportunity [180], PAMAP2 [177], Skoda [245]	F1 score	About 72.6%, 85.4%, 92.4% for respective dataset
[166]	2018	Structural RNN	Daily activity, locomotion	Collective Activity [42], Volleyball [92]	Accuracy	89.1%, 89.3% for respective dataset
[145]	2020	GCN+MLP	Daily activity	EGTEA [120], EPIC-Kitchens [50]	Mean average precision	About 29.4%, 51.6%, respectively
[211]	2019	CNN	Daily activity	THU-READ [210], WCVS [139]	Accuracy	91.72%, 67.04%, respectively
[231]	2020	CNN	Daily activity	EGTEA [120], EPIC-Kitchens [50]	Accuracy	40.5% (Top-5), 62.7%, respectively
[56]	2019	CNN+LSTM	Locomotion activity	Self-collected dataset	Accuracy	94%
[54]	2016	Binarized BLSTM	Daily activity, locomotion	Opportunity [180], PAMAP2 [177]	F1 score	About 76%, 92% for respective dataset
[119]	2017	CGAN + VGG-LSTM	Locomotion, health-related activity	Private video data	Accuracy, Precision, Recall, F1 score	84.7%, 76.7% (F1) for respective data
[59]	2018	GAN + LSTM	Daily activity, locomotion	Collective Activity [42], Volleyball [92]	Accuracy	91.7%, 93% for respective dataset
[192]	2018	CNN + AE	Locomotion activity	Private data collected by Radar	Accuracy	94.2%
[264]	2018	AE + CNN + LSTM	Locomotion activity	Private CSI data	Accuracy, True Positive Rate, False Positive Rate	97.4% (Accuracy)
[131]	2019	CNN + GRU	Locomotion activity factory maintenance	STISEN [203], Skoda [245], PAMAP2 [177]	F1 score	About 96.5%, 93.1%, 89.3% for respective dataset
[33]	2019	LSTM + Attention Model	Locomotion activity	MHEALTH [16], PAMAP2 [177], UCI HAR [10]	Accuracy, Precision, Recall, F1 score	About 96.1%, 89.9%, 85.5% (F1) for respective dataset

advantageous of its simplicity and speed, and is often applied in the cases where there is a large dataset available or the training of the model is very slow. Nevertheless, it has high variability, since the difference in the training data and test data can lead to a significant difference in the classification accuracy [144].

Table 4. Confusion Matrix of a Binary Classification Task

	Actual Positive	**Actual Negative**
Predicted Positive	True positive (tp)	False positive (fp)
Predicted Negative	False negative (fn)	True negative (tn)

In practice, it is common to repeat the evaluation for multiple times (e.g., 30), since deep models are stochastic and involve different sources of randomness (e.g., random initial weights) [26]. In other words, the same model is evaluated on the same data for multiple times and only change the seed to generate random numbers. Then, the average performance can be taken as the performance of the model.

6.1.2 Cross Validation. Cross validation is a commonly used evaluation method. The most popular cross validation is k-fold cross validation, which divides the dataset into k sets where one set is used as test data and the remaining $k - 1$ sets as training data. This process is repeated k times with each of these sets used exactly once as the test data. The value of k is usually 5 or 10. The average performance over all k trials is used as the performance of the model. In addition to k-fold cross validation, there are other cross validation methods such as leave-one-out, and shuffle-split. Compared to the holdout method, cross validation is more reliable and can use data more effectively [142]. The main disadvantage of cross validation is that it increases the computational cost. Cross validation is applicable for small datasets.

In the evaluation of deep models, it is also recommended to repeat the cross validation process for multiple times with only the change of random seed.

6.1.3 Bootstrap. Bootstrap is another method that can be used to evaluate a deep model, which involves resampling a dataset with replacement many times to obtain the statistical performance and corresponding confidence. It performs better for small datasets than cross validation [102]. It has been used in deep learning for different tasks [82, 174] though it is still not very popular.

6.2 Evaluation Metrics

There are many metrics that can be used to measure the classification performance of deep models [88, 165, 205], and the commonly used ones include *accuracy*, *error rate*, *precision*, *recall*, *F-measure*, **Receiver Operating Characteristic (ROC)** curve, and **Area Under the Curve (AOC)**. For clarification, we introduce these metrics using the confusion matrix of a binary classification task to predict the presence of human in an image (presence is encoded as 1, and non-presence as 0), but these metrics are also applicable for multi-class classification tasks. Table 4 shows the confusion matrix, which we will use to explain different metrics. The row of the table is the predicted label, and the column is the actual label. A true positive (denoted by tp) means that the predicted label is 1 and the actual label of the data example is 1, and a true negative (tn) means that the predicted label is 0, and the actual label is 0. A false positive (fp) represents that the predicted label is 1, but the actual label is 0. A false negative (fn) represents that the predicted label is 0, but the actual label is 0.

6.2.1 Accuracy and Error Rate. A commonly used classification performance metric is *accuracy*, which measures the proportion of correct predictions by the total number of data examples. It is written as follows:

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}. \tag{3}$$

Another metric related to *accuracy* is *error rate*, which measures the proportion of incorrect predictions by the total number of data examples. It is described as follows:

$$\begin{aligned} error\_rate &= \frac{fp + fn}{tp + fp + tn + fn} \\ &= 1 - accuracy. \end{aligned} \tag{4}$$

Note that *accuracy* and *error rate* are not suitable for the case where the data are very imbalanced.

6.2.2 Precision and Recall. *Precision* and *recall* are another two classification metrics that are usually used together. *Precision* measures the proportion of correctly predicted positive cases (human presence) by the total number of predicted positive cases, namely

$$precision = \frac{tp}{tp + fp}. \tag{5}$$

Recall measures the proportion of the correctly predicted positive cases by the total correct predicted cases, namely

$$recall = \frac{tp}{tp + tn}. \tag{6}$$

The *precision* provides information about the performance of a model regarding false positives, while the *recall* provides information about the performance regarding false negatives [205].

6.2.3 F-measure. Using both *precision* and *recall* metrics to evaluate the effect of each parameter of the model might be troublesome. One way to address this trouble is to use the harmonic mean of the two metrics, which is called *F-measure*. It is written as

$$F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{7}$$

6.2.4 True Positive Rate (Sensitivity). The true positive rate (tpr), also known as *sensitivity*, measures the proportion of correctly classified positive cases by the number of actual positive cases, namely

$$tpr = \frac{tp}{tp + fn}. \tag{8}$$

6.2.5 False Positive Rate. The false positive rate (fpr), also known as *fall-out*, measures the proportion of the number of negative cases that are wrongly classified as positive by the number of actual negative cases, namely

$$fpr = \frac{fp}{fp + tn}. \tag{9}$$

6.2.6 ROC. A ROC graph is used to visualize the performance of a classifier [58], which shows the relationship between true positive rate and false positive rate. It is believed to provide a richer information than scalar measures such as accuracy or error rate. Figure 8 shows a basic ROC graph with three classifiers. The closer a classifier plotted to the upper left corner, the better performance it has. In Figure 8, classifier 1 performs better than classifiers 2 and 3. Note that the performance of classifier 2 is actually equal to random guess (dashed line in Figure 8).

Fawcett [58] suggests that a classifier is more conservative if it is closer to the left-hand side of the graph; whereas it is more liberal if it is closer to the upper right corner. The concept is similar to the precision-recall tradeoff. Being conservative means minimizing false positives at a cost of missing some true positives; whereas being liberal means maximizing true positives at a cost of including more false positives.

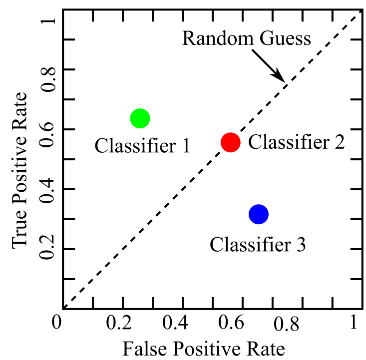


Fig. 8. A simple ROC graph with three binary classifiers.

6.2.7 AUC. AUC is another performance metric, which measures the area under the ROC curve. Compared to the ROC curve, which depicts the classification performance in 2D, AUC is a single scalar value [58]. The value of AUC ranges between 0 and 1, and random guessing has an area of 0.5. The greater the AUC value, the better the classification performance.

6.3 Datasets

There are a lot of public datasets available for HAR. We summarize the statistics of the common datasets in Table 5, including year of data collected, sensor used, activity type, number of activity, number of subject, number of instances, and download link. Most datasets are collected using inertial sensors and cameras, but there are some relatively new datasets that are collected using event cameras. More datasets that contain multi-modal signals for more types of activity would be helpful in promoting the development and advances of HAR with deep learning.

7 CONCLUSION AND OPEN CHALLENGES

In this article, we provide a comprehensive tutorial on deep learning methods for HAR. In particular, we introduce the related techniques in detail, ranging from preprocessing, model building, to evaluation. While this study is mainly developed in the context of HAR, the related techniques can be used for other tasks such as image processing, and speech recognition.

In the following, we provide open challenges that remain in deep learning or HAR.

• **Multimodal and universal activity recognition**. Activity has different levels. While it is easy to recognize simple activity (e.g., walking, running) using single sensor (e.g., accelerometer), complex activities (e.g., in a meeting) require to combine multiple sensors for accurate recognition. In recent years, some studies (e.g., References [70, 260]) have demonstrated that the combination of multiple sensors helps improve the classification accuracy of locomotion activities. However, most existing studies focus on recognizing few activities. Methods for recognizing more universal activities are required to be further explored based on multimodal signals from both ambient sensors and mobile sensors.

• **Developing new deep models that require less labeled data**. Collecting labeled data is usually expensive, which requires plenty of effort and time. Generative deep models (e.g., AEs and GANs) are able to make use of unsupervised data, but they are not directly applicable for HAR. New deep models that can be trained with few labeled data need to be developed. Hybrid models [192, 264], which combine generative models with discriminative models, are promising. Several

Table 5. Public Datasets for HAR

Dataset	Year	Sensor	Activity Type	#Activity	#Subject	#Instances	Download Link
Kasteren [218]	2008	Wireless sensor (RFM DM 1810)	Daily activity	8	1	245	Website
Skoda Mini Checkpoint [245]	2008	Inertial sensor	Factory maintenance	10	1	700	Website
UC Berkeley WARD [240]	2009	Inertial sensors	Locomotion acitivity	13	20	N/A	Website
Daphnet Freezing of Gait Syndrome [15]	2009	Inertial sensors	Health-related activity	3	10	237	Website
CMU-MMAC [51]	2009	Inertial sensors, camera, microphone, and so on	Daily activity (kitchen)	5	5	N/A	Website
Opportunity [180]	2010	Motion sensors	Daily activity	9	4	2,551	Website
HMDB-51 [107]	2011	Camera	Locomotion activity	51	N/A	6,849	Website
WISDM Actitracker Dataset [108]	2011	Accelerometer	Locomotion activity	6	29	5, 424	Website
ADL [10]	2012	Inertial sensors	Locomotion activity	6	30	10, 299	Website
UCF-101 [202]	2012	Camera	Locomotion activity	101	N/A	13,320	Website
USC-HAD [249]	2012	Inertial sensors	Locomotion activity	12	14	N/A	Website
PAMAP2 [177]	2012	Inertial sensors, heart rate monitor	Daily activity	18	9	N/A	Website
Berkeley MHAD [153]	2013	Camera, inertial sensors, microphone, and so on	Daily activity	11	12	660	Website
REALDISP [17]	2014	Inertial sensors	Locomotion, daily activity	33	17	1,419	Website
ActRecTut [28]	2014	Inertial sensors	Gesture	11	2	337	Website
HHAR [203]	2015	Inertial sensors	Locomotion activity	6	9	N/A	Website
ActivityNet-200 [29]	2016	Camera	Daily activity, entertainment, and so on	200	N/A	27, 801	Website
AReM [156]	2016	WiFi, accelerometer	Locomotion activity	6	N/A	42,240	Website
UniMiB SHAR [135]	2016	Accelerometer	Locomotion activity	9	30	11, 771	Website
UMAFall [32]	2016	Wearable sensors, inertial sensors	Daily activity	17	19	531	Website
UCF-50 DVS [89]	2016	Event camera	Daily activity	50	N/A	6,676	Website
DVS128 Gesture [8]	2017	Event camera	Gesture	11	29	1, 342	Website
Wifi-Activity-Recognition [244]	2017	WiFi (CSI)	Locomotion activity	6	N/A	720	Website
Orange4Home [48]	2017	236 heterogeneous sensors	Routine daily activity	20	1	493	Website
Kinetics-400 [100]	2017	Camera	daily activity, locomotion activity, entertainment, and so on	400	N/A	306,245	Website
Kinetics-700 [31]	2019	Camera	daily activity, locomotion activity, entertainment, and so on	700	N/A	650,317	Website
Epic Kitchen [50]	2018	Head-mounted camera	Kitchen activity	125	32	39,596	Website
LAR [70]	2018	Inertial sensors, barometer	Locomotion activity	8	12	12,822	Website
WESAD [187]	2018	ECG, EDA, EMG, respiration, body temperature, accelerometer	Emotion	15	N/A	63,000,000	Website
Multi-site Sensing [209]	2019	Accelerometer	Locomotion activity	22	42	6,674	Website
DVS Action [134]	2019	Event camera	Gesture	10	15	450	Website
DVS Fall [134]	2019	Event camera	Locomotion activity	4	15	180	Website

studies have been done though they are still in their infancy. It is also promising to develop new semi-supervised deep models [38] and active deep models [214] that require less labeled data.

• **Crowdsourcing quality data for deep models**. Deep models generally require to be trained on a large scale dataset; however, collecting such dataset can be cost prohibitive. A possible way to reduce the cost is to use crowdsourcing that allows each individual to contribute their own data. However, the challenge of crowdsourcing data is how to guarantee the quality of collected data. It worths to investigate or develop strategies to encourage individuals to share their data and to ensure the quality of data.

• **Efficient deep learning algorithms for resource limited devices (e.g., smartphones)**. Deep models are often computationally expensive, and most of them require to run on a server or PC with high configuration. Smart devices (e.g., smartphones, smart watches) have become prevalent in modern life these days, but their computing capability, though has been significantly improved, is still insufficient to run most deep models. More efficient deep models should be developed to be run on mobile devices [53], which will bring great convenience for people's life.

• **Stable and robust deep models**. While deep models have been shown effective and superior in many tasks, their performance may not be stable when the training data suffering small perturbations. This can reduce users' trust especially in health-related applications. Although some attempts have been done [76, 77], more works on developing stable and robust models are needed.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/.

[2] Heba Abdelnasser, Moustafa Youssef, and Khaled A. Harras. 2015. Wigest: A ubiquitous wifi-based gesture recognition system. In *Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM'15).* IEEE, 1472–1480.

[3] Jake K. Aggarwal and Michael S. Ryoo. 2011. Human activity analysis: A review. *ACM Comput. Surv.* 43, 3 (2011), 16.

[4] Unaiza Ahsan, Chen Sun, and Irfan Essa. 2017. DiscrimNet: Semi-supervised action recognition from videos using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops .Women in Computer Vision (WiCV'17).*

[5] Bandar Almaslukh, Jalal AlMuhtadi, and Abdelmonim Artoli. 2017. An effective deep autoencoder approach for online smartphone-based human activity recognition. *Int. J. Comput. Sci. Netw. Secur.* 17, 4 (2017), 160–165.

[6] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. 2016. Deep activity recognition models with triaxial accelerometers. In *Proceedings of the Workshops at the 30th AAAI Conference on Artificial Intelligence.*

[7] Kerem Altun, Billur Barshan, and Orkun Tunçel. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recogn.* 43, 10 (2010), 3605–3620.

[8] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. 2017. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7243–7252.

[9] Ian Anderson and Henk Muller. 2006. Practical activity recognition using gsm data. In *Proceedings of the 5th International Semantic Web Conference (ISWC'06)*, Vol. 1.

[10] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'13).*

[11] Ery Arias-Castro, David L. Donoho, et al. 2009. Does median filtering truly preserve edges better than linear filtering? *Ann. Stat.* 37, 3 (2009), 1172–1206.

[12] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. arXiv:1701.07875. Retrieved from https://arxiv.org/abs/1701.07875.

[13] Martin Arjovsky, Amar Shah, and Yoshua Bengio. 2016. Unitary evolution recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*. 1120–1128.

[14] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M. Hausdorff, Nir Giladi, and Gerhard Troster. 2009. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Trans. Inf. Technol. Biomed.* 14, 2 (2009), 436–446.

[15] Marc Bachlin, Daniel Roggen, Gerhard Troster, Meir Plotnik, Noit Inbar, Inbal Meidan, Talia Herman, Marina Brozgol, Eliya Shaviv, Nir Giladi, et al. 2009. Potentials of enhanced context awareness in wearable assistants for Parkinson's disease patients with the freezing of gait syndrome. In *Proceedings of the International Symposium on Wearable Computers.* IEEE, 123–130.

[16] Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: A novel framework for agile development of mobile health applications. In *Proceedings of the International Workshop on Ambient Assisted Living*. Springer, 91–98.

[17] Oresti Banos, Mate Toth, Miguel Damas, Hector Pomares, and Ignacio Rojas. 2014. Dealing with the effects of sensor displacement in wearable activity recognition. *Sensors* 14, 6 (2014), 9995–10023.

[18] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 65–74.

[19] Yoshua Bengio et al. 2009. Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1 (2009), 1–127.

[20] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. 2013. Advances in optimizing recurrent networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8624–8628.

[21] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828.

[22] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: A CPU and GPU math compiler in Python. In *Proceedings of the 9th Python in Science Conference*, Vol. 1. 3–10.

[23] David Berthelot, Thomas Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. arXiv:1703.10717. Retrieved from https://arxiv.org/abs/1703.10717.

[24] Sourav Bhattacharya and Nicholas D. Lane. 2016. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops'16)*. IEEE, 1–6.

[25] Ronald Newbold Bracewell and Ronald N. Bracewell. 1986. *The Fourier Transform and its Applications*. Vol. 31999. McGraw–Hill New York.

[26] Jason Brownlee. 2019. How to Evaluate the Skill of Deep Learning Models. Retrieved from https://machinelearningmastery.com/evaluate-skill-deep-learning-models/.

[27] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* 46, 3 (2014), 33.

[28] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *Comput. Surv.* 46, 3 (2014), 33:1–33:33. https://doi.org/10.1145/2499621

[29] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[30] Tom Campbell. 1981. Seven Theories of Human Society. Clarendon Press, 169–229.

[31] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. arXiv:1907.06987. Retrieved from https://arxiv.org/abs/1907.06987.

[32] Eduardo Casilari, Jose A. Santoyo-Ramón, and Jose M. Cano-García. 2017. Umafall: A multisensor dataset for the research on automatic fall detection. *Proc. Comput. Sci.* 110 (2017), 32–39.

[33] Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. 2019. Multi-agent attentional activity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. 1344–1350.

[34] Liming Chen, Jesse Hoey, Chris D. Nugent, Diane J. Cook, and Zhiwen Yu. 2012. Sensor-based activity recognition. *IEEE Trans. Syst. Man Cyberneti. C* 42, 6 (2012), 790–808.

[35] Ling Chen, Yi Zhang, and Liangying Peng. 2020. METIER: A deep multi-task learning based activity and user recognition model using wearable sensors. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 4, 1 (2020), 1–18.

[36] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *Neural Information Processing Systems, Workshop on Machine Learning Systems*.

[37] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.

[38] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2018. Semi-supervised deep learning with memory. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 268–283.

[39] Zhibo Chen, Wei Zhou, and Weiping Li. 2017. Blind stereoscopic video quality assessment: From depth perception to overall experience. *IEEE Trans. Image Process.* 27, 2 (2017), 721–734.

[40] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

[41] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1724–1734.

[42] Wongun Choi, Khuram Shahid, and Silvio Savarese. 2009. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proceedings of the IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops'09)*. IEEE, 1282–1289.

[43] François Chollet et al. 2015. Keras. Retrieved from https://keras.io.

[44] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. P. Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt. 2008. The mobile sensing platform: An embedded activity recognition system. *IEEE Perv. Comput.* 7, 2 (Apr. 2008), 32–41. https://doi.org/10.1109/MPRV.2008.39

[45] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Workshop on Deep Learning (NIPS'14)*.

[46] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A matlab-like environment for machine learning. In *Proceedings of the Conference on Neural Information Processing Systems BigLearn Workshop (NIPS BigLearn Workshop'11)*.
[47] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2018. Generative adversarial networks: An overview. *IEEE Sign. Process. Mag.* 35, 1 (2018), 53–65.
[48] Julien Cumin, Grégoire Lefebvre, Fano Ramparany, and James L. Crowley. 2017. A dataset of routine daily activities in an instrumented home. In *Proceedings of the International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, 413–425.
[49] Jason Dai, Yiheng Wang, Xin Qiu, Ding Ding, Yao Zhang, Yanzhang Wang, Xianyan Jia, Cherry Zhang, Yan Wan, Zhichao Li, et al. 2018. Bigdl: A distributed deep learning framework for big data. In *Proceedings of the ACM Symposium on Cloud Computing (SoCC'19)*. 50–60.
[50] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*.
[51] Fernando De la Torre, Jessica Hodgins, Javier Montano, Sergio Valcarcel, R. Forcada, and J. Macey. 2008. *Guide to the Carnegie Mellon University Multimodal Activity (cmu-mmac) Database*. Technical Report. Carnegie Mellon University, Pittsburgh, PA.
[52] Li Deng, Dong Yu, et al. 2014. Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing* 7, 3–4 (2014), 197–387.
[53] Yunbin Deng. 2019. Deep learning on mobile devices: A review. In *Proc. SPIE 10993, Mobile Multimedia/Image Processing, Security, and Applications*. May 2019. Art. no. 109930.
[54] Marcus Edel and Enrico Köppe. 2016. Binarized-blstm-rnn based human activity recognition. In *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN'16)*. IEEE, 1–7.
[55] Shahla Faisal and Gerhard Tutz. 2017. Missing value imputation for gene expression data by tailored nearest neighbors. *Stat. Appl. Genet. Molec. Biol.* 16, 2 (2017), 95–106.
[56] Xiaoyi Fan, Fangxin Wang, Feng Wang, Wei Gong, and Jiangchuan Liu. 2019. When rfid meets deep learning: Exploring cognitive intelligence for activity identification. *IEEE Wireless Commun.* 26, 3 (2019), 19–25.
[57] Hongqing Fang and Chen Hu. 2014. Recognizing human activity in smart home using deep learning algorithm. In *Proceedings of the 33rd Chinese Control Conference*. IEEE, 4716–4720.
[58] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 8 (2006), 861–874.
[59] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2018. Multi-level sequence GAN for group activity recognition. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 331–346.
[60] Salvador García, Julián Luengo, and Francisco Herrera. 2015. *Data Preprocessing in Data Mining*. Springer.
[61] Enrique Garcia-Ceja and Ramon Brena. 2013. Long-term activity recognition from accelerometer data. *Proc. Technol.* 7 (2013), 248–256.
[62] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.
[63] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220.
[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
[65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.
[66] Alex Graves. 2013. Generating sequences with recurrent neural networks. arXiv:1308.0850. Retrieved from https://arxiv.org/abs/1308.0850.
[67] Fuqiang Gu, Xuke Hu, Milad Ramezani, Debaditya Acharya, Kourosh Khoshelham, Shahrokh Valaee, and Jianga Shang. 2019. Indoor localization improved by spatial context—A survey. *Comput. Surv.* 52, 3, Article 64 (Jul. 2019), 35 pages. https://doi.org/10.1145/3322241
[68] Fuqiang Gu, Allison Kealy, Kourosh Khoshelham, and Jianga Shang. 2015. User-independent motion state recognition using smartphone sensors. *Sensors* 15, 12 (2015), 30636–30652.
[69] Fuqiang Gu, Kourosh Khoshelham, and Shahrokh Valaee. 2017. Locomotion activity recognition: A deep learning approach. In *Proceedings of the IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC'17)*. IEEE, 1–5.
[70] Fuqiang Gu, Kourosh Khoshelham, Shahrokh Valaee, Jianga Shang, and Rui Zhang. 2018. Locomotion activity recognition using stacked denoising autoencoders. *IEEE IoT J.* 5, 3 (2018), 2085–2093.
[71] Fuqiang Gu, Kourosh Khoshelham, Chunyang Yu, and Jianga Shang. 2018. Accurate step length estimation for pedestrian dead reckoning localization using stacked autoencoders. *IEEE Trans. Instrum. Meas.* 68, 8 (2018), 2705–2713.

[72] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. 2018. Recent advances in convolutional neural networks. *Pattern Recogn.* 77 (2018), 354–377.

[73] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* 1, 2 (2017), 1–28.

[74] Guodong Guo and Alice Lai. 2014. A survey on still image based human action recognition. *Pattern Recogn.* 47, 10 (2014), 3343–3361.

[75] Sojeong Ha and Seungjin Choi. 2016. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'16).* IEEE, 381–388.

[76] Eldad Haber, Keegan Lensink, Eran Triester, and Lars Ruthotto. 2019. IMEXnet: Aforward stable deep neural network. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19).* PMLR 97:2525–2534.

[77] Eldad Haber and Lars Ruthotto. 2017. Stable architectures for deep neural networks. *Inverse Probl.* 34, 1 (2017), 014004.

[78] Mahmudul Hasan and Amit K. Roy-Chowdhury. 2015. A continuous learning framework for activity recognition using deep hybrid feature models. *IEEE Trans. Multimedia* 17, 11 (2015), 1909–1922.

[79] Mohammed Mehedi Hassan, Md Zia Uddin, Amr Mohamed, and Ahmad Almogren. 2018. A robust human activity recognition system using smartphone sensors and deep learning. *Fut. Gener. Comput. Syst.* 81 (2018), 307–313.

[80] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[81] Marko Helen and Tuomas Virtanen. 2005. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of the 13th European Signal Processing Conference.* IEEE, 1–4.

[82] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.*

[83] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*, Vol. 3.

[84] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Sign. Process. Mag.* 29 (2012).

[85] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 7 (2006), 1527–1554.

[86] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.

[87] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[88] Mohammad Hossin and M. N. Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manage. Process* 5, 2 (2015), 1.

[89] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. 2016. DVS benchmark datasets for object tracking, action recognition, and object recognition. *Front. Neurosci.* 10 (2016), 405.

[90] Dana Hughes and Nikolaus Correll. 2018. Distributed convolutional neural networks for human activity recognition in wearable robotics. In *Distributed Autonomous Robotic Systems.* Springer, 619–631.

[91] Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'08)*, Vol. 8., 10–19.

[92] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1971–1980.

[93] Ozlem Durmaz Incel, Mustafa Kose, and Cem Ersoy. 2013. A review and taxonomy of activity recognition on mobile phones. *BioNanoScience* 3, 2 (2013), 145–171.

[94] Masaya Inoue, Sozo Inoue, and Takeshi Nishida. 2018. Deep recurrent neural network for mobile human activity recognition with high throughput. *Artif. Life Robot.* 23, 2 (2018), 173–185.

[95] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia.* ACM, 675–678.

[96] Li Jing, Caglar Gulcehre, John Peurifoy, Yichen Shen, Max Tegmark, Marin Soljacic, and Yoshua Bengio. 2019. Gated orthogonal recurrent units: On learning to forget. *Neural Comput.* 31, 4 (2019), 765–783.

[97] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality. In *Sixth International Conference on Learning Representations (ICLR'18).*

[98] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4401–4410.

[99] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi, and Nobuhiko Nishio. 2011. HASC Challenge: Gathering large scale human activity corpus for the real-world activity understandings. In *Proceedings of the 2nd Augmented Human International Conference.* ACM, 27.

[100] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. arXiv:1705.06950. Retrieved from https://arxiv.org/abs/1705.06950.

[101] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. Optimal whitening and decorrelation. *Am. Stat.* 72, 4 (2018), 309–314.

[102] Ji-Hyun Kim. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* 53, 11 (2009), 3735–3745.

[103] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR'13).*

[104] Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'11).* IEEE, 3241–3248.

[105] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images.* Master's thesis. Department of Computer Science, University of Toronto.

[106] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems.* 1097–1105.

[107] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision.* IEEE, 2556–2563.

[108] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explor. Newslett.* 12, 2 (2011), 74–82.

[109] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 283–294.

[110] Oscar D. Lara and Miguel A. Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* 15, 3 (2012), 1192–1209.

[111] Quoc V. Le, Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pang W. Koh, and Andrew Y. Ng. 2010. Tiled convolutional neural networks. In *Advances in Neural Information Processing Systems.* 1279–1287.

[112] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

[113] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[114] Chen-Yu Lee, Patrick Gallagher, and Zhuowen Tu. 2017. Generalizing pooling functions in cnns: Mixed, gated, and tree. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2017), 863–875.

[115] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning.* ACM, 609–616.

[116] Dan Li, Jitender Deogun, William Spaulding, and Bill Shuart. 2004. Towards missing data imputation: a study of fuzzy k-means clustering method. In *Proceedings of the International Conference on Rough Sets and Current Trends in Computing.* Springer, 573–579.

[117] Fei-Fei Li, Justin Johnson, and Serena Yeung. 2019. CS231n Convolutional Neural Networks for Visual Recognition. Retrieved from http://cs231n.github.io/neural-networks-2/.

[118] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S. Burd. 2016. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM.* 164–175.

[119] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Yueyang Chen, Huangcan Li, Ivan Marsic, and Randall S. Burd. 2017. Region-based activity recognition using conditional GAN. In *Proceedings of the 25th ACM International Conference on Multimedia.* ACM, 1059–1067.

[120] Yin Li, Miao Liu, and James M. Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV'18).* 619–635.

[121] Yongmou Li, Dianxi Shi, Bo Ding, and Dongbo Liu. 2014. Unsupervised feature learning for human activity recognition using smartphone sensors. In *Mining Intelligence and Knowledge Exploration.* Springer, 99–107.

[122] Zuhe Li, Yangyu Fan, and Weihua Liu. 2015. The effect of whitening transformation on pooling operations in convolutional autoencoders. *EURASIP J. Adv. Sign. Process.* 2015, 1 (2015), 37.

[123] Ming Liang and Xiaolin Hu. 2015. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3367–3375.

[124] Lin Liao, Dieter Fox, and Henry A. Kautz. 2005. Location-Based Activity Recognition using Relational Markov Networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'05)*, Vol. 5. 773–778.

[125] Kang Ling, Haipeng Dai, Yuntang Liu, Alex X. Liu, Wei Wang, and Qing Gu. 2020. Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Trans. Mobile Comput.* (2020).

[126] Si Liu, Yao Sun, Defa Zhu, Renda Bao, Wei Wang, Xiangbo Shu, and Shuicheng Yan. 2017. Face aging with contextual generative adversarial nets. In *Proceedings of the 25th ACM International Conference on Multimedia.* ACM, 82–90.

[127] Yunhao Liu, Yiyang Zhao, Lei Chen, Jian Pei, and Jinsong Han. 2011. Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays. *IEEE Trans. Parallel Distrib. Syst.* 23, 11 (2011), 2138–2149.

[128] Hong Lu, A. J. Bernheim Brush, Bodhi Priyantha, Amy K. Karlson, and Jie Liu. 2011. Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. In *Proceedings of the International Conference on Pervasive Computing.* Springer, 188–205.

[129] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the ACM Conference on Ubiquitous Computing.* ACM, 351–360.

[130] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems.* ACM, 71–84.

[131] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: multi-level attention mechanism for multimodal human activity recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence.* AAAI Press, 3109–3115.

[132] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision.* 2794–2802.

[133] Mathworks. 2016. Deep Learning Toolbox. Retrieved from https://www.mathworks.com/products/deep-learning.html.

[134] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. 2019. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Front. Neurorobot.* 13 (2019), 38.

[135] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2017. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Appl. Sci.* 7, 10 (2017), 1101.

[136] Tomas Mikolov, Armand Joulin, Sumit Chopra, Michael Mathieu, and Marc'Aurelio Ranzato. 2014. Learning longer memory in recurrent neural networks. arXiv:1412.7753. Retrieved from https://arxiv.org/abs/1412.7753.

[137] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv:1411.1784. Retrieved from https://arxiv.org/abs/1411.1784.

[138] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. 2006. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Understand.* 104, 2–3 (2006), 90–126.

[139] Mohammad Moghimi, Pablo Azagra, Luis Montesano, Ana C. Murillo, and Serge Belongie. 2014. Experiments on an rgb-d wearable vision system for egocentric activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 597–603.

[140] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. 2009. Deep belief networks for phone recognition. In *Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, Vol. 1. Vancouver, Canada, 39.

[141] Subhas Chandra Mukhopadhyay. 2014. Wearable sensors for human activity monitoring: A review. *IEEE Sens. J.* 15, 3 (2014), 1321–1330.

[142] Andreas C. Müller, Sarah Guido, et al. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, Inc.

[143] Abdulmajid Murad and Jae-Young Pyun. 2017. Deep recurrent neural networks for human activity recognition. *Sensors* 17, 11 (2017), 2556.

[144] Steve Mutuvi. 2019. Introduction to Machine Learning Model Evaluation. Retrieved from https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f.

[145] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. 2020. EGO-TOPO: Environment Affordances from Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 163–172.

[146] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. 2014. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems.* 424–432.

[147] Andrew Ng. 2019. UFLDL Tutorial: PCA Whitening. Retrieved from http://ufldl.stanford.edu/tutorial/unsupervised/PCAWhitening/.
[148] Andrew Ng et al. 2011. Sparse autoencoder. *CS294A Lect. Not.* 72, 2011 (2011), 1–19.
[149] Trung Thanh Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. 2015. Similar gait action recognition using an inertial sensor. *Pattern Recogn.* 48, 4 (2015), 1289–1301.
[150] Kai Niu, Fusang Zhang, Zhaoxin Chang, and Daqing Zhang. 2018. A Fresnel Diffraction Model Based Human Respiration Detection System Using COTS Wi-Fi Devices. In *Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers.* ACM, 416–419.
[151] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. 2018. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Syst. Appl.* 105 (2018), 233–261.
[152] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* 2642–2651.
[153] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'13).* IEEE, 53–60.
[154] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, et al. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'11).* IEEE, 3153–3160.
[155] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quéot. 2013. TRECVID 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. TRECVID Publications. Retrieved August 22, 2021 http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.
[156] Filippo Palumbo, Claudio Gallicchio, Rita Pucci, and Alessio Micheli. 2016. Human activity recognition using multisensor data fusion based on reservoir computing. *J. Ambient Intell. Smart Environ.* 8, 2 (2016), 87–107.
[157] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. 2019. Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access* 7 (2019), 36322–36333.
[158] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*
[159] Ling Pei, Jingbin Liu, Robert Guinness, Yuwei Chen, Heidi Kuusniemi, and Ruizhi Chen. 2012. Using LS-SVM based motion recognition for smartphone indoor wireless positioning. *Sensors* 12, 5 (2012), 6155–6175.
[160] Ling Pei, Songpengcheng Xia, Lei Chu, Fanyi Xiao, Qi Wu, Wenxian Yu, and Robert Qiu. 2021. MARS: Mixed Virtual and Real Wearable Sensors for Human Activity Recognition with Multi-Domain Deep Learning Model. *IEEE IoT J.* (2021).
[161] Cuong Pham and Patrick Olivier. 2009. Slice&dice: Recognizing food preparation activities using embedded accelerometers. In *Proceedings of the European Conference on Ambient Intelligence.* Springer, 34–43.
[162] Thomas Plötz, Nils Y. Hammerla, and Patrick L. Olivier. 2011. Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the 22d International Joint Conference on Artificial Intelligence.*
[163] Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2537–2544.
[164] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 6 (2010), 976–990.
[165] David Martin Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.
[166] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. 2018. stagnet: An attentive semantic RNN for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV'18).* 101–117.
[167] Xin Qi, Gang Zhou, Yantao Li, and Ge Peng. 2012. Radiosense: Exploiting wireless communication patterns for body sensor network activity recognition. In *Proceedings of the IEEE 33rd Real-Time Systems Symposium.* IEEE, 95–104.
[168] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing.* ACM, 281–290.
[169] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434. Retrieved from https://arxiv.org/abs/1511.06434.
[170] Valentin Radu and Maximilian Henne. 2019. Vision2Sensor: Knowledge transfer across sensing modalities for human activity recognition. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 3, 3 (2019), 84.

[171] Valentin Radu, Panagiota Katsikouli, Rik Sarkar, and Mahesh K. Marina. 2014. A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. ACM, 280–294.

[172] Valentin Radu, Nicholas D. Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2016. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 185–188.

[173] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 1, 4 (2018), 157.

[174] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv:1711.05225. Retrieved from https://arxiv.org/abs/1711.05225.

[175] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity recognition from accelerometer data. In *Proceedings of the AAAI Annual Conference on Artificial Intelligence (AAAI'05)*, Vol. 5. 1541–1546.

[176] Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* 24, 5 (2013), 971–981.

[177] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the 16th International Symposium on Wearable Computers*. IEEE, 108–109.

[178] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[179] Malcolm Reynolds, Gabriel Barth-Maron, Frederic Besse, Diego de Las Casas, Andreas Fidjeland, Tim Green, Adrià Puigdomènech, Sébastien Racanière, Jack Rae, and Fabio Viola. 2017. Open Sourcing Sonnet—A New Library for Constructing Neural Networks. Retrieved from https://deepmind.com/blog/open-sourcing-sonnet/.

[180] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *Proceedings of the 7th International Conference on Networked Sensing Systems (INSS'10)*. IEEE, 233–240.

[181] Charissa Ann Ronao and Sung-Bae Cho. 2015. Deep convolutional neural networks for human activity recognition with smartphone sensors. In *Proceedings of the International Conference on Neural Information Processing*. Springer, 46–53.

[182] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* 59 (2016), 235–244.

[183] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323 (1986), 533–536.

[184] Ruslan Salakhutdinov. 2015. Learning deep generative models. *Annu. Rev. Stat. Appl.* 2 (2015), 361–385.

[185] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Deep boltzmann machines. In *Artificial Intelligence and Statistics*. 448–455.

[186] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. arXiv:1801.01078. Retrieved from https://arxiv.org/abs/1801.01078.

[187] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 400–408.

[188] Markus Scholz, Till Riedel, Mario Hock, and Michael Beigl. 2013. Device-free and device-bound activity recognition using radio signal strength. In *Proceedings of the 4th Augmented Human International Conference*. ACM, 100–107.

[189] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, Vol. 3. IEEE, 32–36.

[190] Frank Seide and Amit Agarwal. 2016. CNTK: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2135–2135.

[191] Ervin Sejdic, Igor Djurovic, et al. 2008. Quantitative performance analysis of scalogram as instantaneous frequency estimator. *IEEE Trans. Sign. Process.* 56, 8 (2008), 3837–3845.

[192] Mehmet Saygın Seyfioğlu, Ahmet Murat Özbayoğlu, and Sevgi Zubeyde Gürbüz. 2018. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. *IEEE Trans. Aerospace Electron. Syst.* 54, 4 (2018), 1709–1723.

[193] Jianga Shang, Fuqiang Gu, Xuke Hu, and Allison Kealy. 2015. Apfiloc: An infrastructure-free indoor localization method fusing smartphone inertial sensors, landmarks and map information. *Sensors* 15, 10 (2015), 27251–27272.

[194] Shuyu Shi, Stephan Sigg, and Yusheng Ji. 2012. Passive detection of situations from ambient fm-radio signals. In *Proceedings of the ACM Conference on Ubiquitous Computing*. ACM, 1049–1053.

[195] Shuyu Shi, Stephan Sigg, Wei Zhao, and Yusheng Ji. 2014. Monitoring attention using ambient FM radio signals. *IEEE Perv. Comput.* 13, 1 (2014), 30–36.

[196] Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga. 2015. A survey of online activity recognition using mobile phones. *Sensors* 15, 1 (2015), 2059–2085.

[197] Stephan Sigg, Ulf Blanke, and Gerhard Tröster. 2014. The telepathic phone: Frictionless activity recognition from wifi-rssi. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PERCOM'14)*. IEEE, 148–155.

[198] Jae Mun Sim, Yonnim Lee, and Ohbyung Kwon. 2015. Acoustic sensor based recognition of human activity in everyday life for smart home services. *Int. J. Distrib. Sens. Netw.* 11, 9 (2015), 679123.

[199] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR'15)*.

[200] Timothy Sohn, Alex Varshavsky, Anthony LaMarca, Mike Y. Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G. Griswold, and Eyal De Lara. 2006. Mobility detection using everyday GSM traces. In *Proceedings of the International Conference on Ubiquitous Computing*. Springer, 212–224.

[201] Cuiling Lan Junliang Xing Wenjun Zeng Song, Sijie and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *The Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 4263–4270.

[202] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012.

[203] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 127–140.

[204] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 591–605.

[205] Mohammed Sunasra. 2019. Performance Metrics for Classification problems in Machine Learning. Retrieved from https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

[206] Ilya Sutskever, James Martens, and Geoffrey E. Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 1017–1024.

[207] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

[208] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[209] Qu Tang, Dinesh John, Binod Thapa-Chhetry, Diego Jose Arguello, and Stephen Intille. 2020. Posture and physical activity detection: Impact of number of sensors and feature type. *Med. Sci. Sports and Exercise* (2020).

[210] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2017. Action recognition in rgb-d egocentric videos. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'17)*. IEEE, 3410–3414.

[211] Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2018. Multi-stream deep neural networks for rgb-d egocentric action recognition. *IEEE Trans. Circ. Syst. Vid. Technol.* 29, 10 (2018), 3001–3015.

[212] Eclipse Deeplearning4j Development Team. Deeplearning4j: Open-source Distributed Deep Learning for the jvm. Retreived from http://deeplearning4j.org.

[213] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: A next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in the 29th Annual Conference on Neural Information Processing Systems (NIPS'15)*, Vol. 5. 1–6.

[214] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. 2019. Bayesian generative active deep learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*.

[215] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (2001), 520–525.

[216] Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. In *Third workshop on Bayesian Deep Learning (NeurIPS'18)*.

[217] Md Zia Uddin, Mohammad Mehedi Hassan, Ahmad Almogren, Atif Alamri, Majed Alrubaian, and Giancarlo Fortino. 2017. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access* 5 (2017), 4525–4536.

[218] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. 2008. Accurate activity recognition in a home setting. In *Proceedings of the 10th International Conference on Ubiquitous Computing*. ACM, 1–9.

[219] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, (Dec. 2010), 3371–3408.

[220] Kishor H. Walse, Rajiv V. Dharaskar, and Vilas M. Thakare. 2016. Pca based optimal ann classifiers for human activity recognition using mobile sensors data. In *Proceedings of 1st International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*. Springer, 429–436.

[221] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. 2019. Joint Activity Recognition and Indoor Localization with WiFi Fingerprints. *IEEE Access* 7 (2019), 80058–80068.

[222] Fangxin Wang, Wei Gong, and Jiangchuan Liu. 2018. On spatial diversity in WiFi-based human activity recognition: A deep learning-based approach. *IEEE IoT J.* 6, 2 (2018), 2035–2047.

[223] Fangxin Wang, Wei Gong, Jiangchuan Liu, and Kui Wu. 2018. Channel selective activity recognition with WiFi: A deep learning approach exploring wideband information. *IEEE Trans. Netw. Sci. Eng.* 7, 1 (2018), 181–192.

[224] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recogn. Lett.* 119 (2019), 3–11.

[225] Jie Wang, Xiao Zhang, Qinhua Gao, Hao Yue, and Hongyu Wang. 2016. Device-free wireless localization and activity recognition: A deep learning approach. *IEEE Trans. Vehic. Technol.* 66, 7 (2016), 6258–6267.

[226] Lukun Wang. 2016. Recognition of human activities using continuous autoencoders with wearable sensors. *Sensors* 16, 2 (2016), 189.

[227] Shuangquan Wang and Gang Zhou. 2015. A review on radio based activity recognition. *Dig. Commun. Netw.* 1, 1 (2015), 20–29.

[228] Weijie Wang, Gaopeng Zhang, Luming Yang, V. S. Balaji, V. Elamaran, and N. Arunkumar. 2019. Revisiting signal processing with spectrogram analysis on EEG, ECG and speech signals. *Fut. Gener. Comput. Syst.* 98 (2019), 227–232.

[229] Wei-zhong Wang, Yan-wei Guo, Bang-yu Huang, Guo-ru Zhao, Bo-qiang Liu, and Lei Wang. 2011. Analysis of filtering methods for 3D acceleration signals in body sensor network. In *Proceedings of the International Symposium on Bioelectronics and Bioinformations*. IEEE, 263–266.

[230] Xuanhan Wang, Lianli Gao, Jingkuan Song, Xiantong Zhen, Nicu Sebe, and Heng Tao Shen. 2018. Deep appearance and motion learning for egocentric activity recognition. *Neurocomputing* 275 (2018), 438–447.

[231] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. 2020. Symbiotic attention with privileged information for ego-centric action recognition. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI'20)*.

[232] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the limit of acoustic gesture recognition. *IEEE Trans. Mobile Comput.* (2020).

[233] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID signal reflection for contact-free activity recognition. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 4 (2018), 1–22.

[234] Zhiguang Wang and Tim Oates. 2015. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Proceedings of the Workshops at the 29th AAAI Conference on Artificial Intelligence*.

[235] Zhengwei Wang, Qi She, and Tomás E. Ward. 2021. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Comput. Surv.* 54, 2, Article 37 (February 2021), 38 pages. https://doi.org/10.1145/3439723

[236] Dan Wu, Daqing Zhang, Chenren Xu, Hao Wang, and Xiang Li. 2017. Device-free WiFi human sensing: From pattern-based to model-based approaches. *IEEE Commun. Mag.* 55, 10 (2017), 91–97.

[237] Fu Xiao, Jing Chen, Xiaohui Xie, Linqing Gui, Lijuan Sun, and Ruchuan Wang. 2018. SEARE: A system for exercise activity recognition and quality evaluation based on green sensing. *IEEE Trans. Emerg. Top. Comput.* 8, 3 (2018), 752–761.

[238] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[239] Yanyang Yan, Wenqi Ren, and Xiaochun Cao. 2018. Recolored image detection via a deep discriminative model. *IEEE Trans. Inf. Forens. Secur.* 14, 1 (2018), 5–17.

[240] Allen Y. Yang, Roozbeh Jafari, S. Shankar Sastry, and Ruzena Bajcsy. 2009. Distributed recognition of human actions using wearable motion sensor networks. *J. Ambient Intell. Smart Environ.* 1, 2 (2009), 103–115.

[241] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*.

[242] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. 2019. Learning gestures from WiFi: A siamese recurrent convolutional architecture. *IEEE IoT J.* 6, 6 (2019), 10763–10772.

[243] Haibo Ye, Tao Gu, Xianping Tao, and Jian Lu. 2016. Scalable floor localization using barometer on smartphone. *Wireless Commun. Mobile Comput.* 16, 16 (2016), 2557–2571.

[244] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. 2017. A survey on behavior recognition using wifi channel state information. *IEEE Commun. Mag.* 55, 10 (2017), 98–104.

[245] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. 2008. Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In *Proceedings of the European Conference on Wireless Sensor Networks.* Springer, 17–33.

[246] Ming Zeng, Le T. Nguyen, Bo Yu, Ole J. Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services.* IEEE, 197–205.

[247] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *the 36th International Conference on Machine Learning (ICML'19).*

[248] Jin Zhang, Fuxiang Wu, Bo Wei, Qieshi Zhang, Hui Huang, Syed W. Shah, and Jun Cheng. 2020. Data augmentation and dense-lstm for human activity recognition using wifi signal. *IEEE IoT J.* (2020).

[249] Mi Zhang and Alexander A. Sawchuk. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing.* ACM, 1036–1043.

[250] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, and David Mcalpine. 2020. A survey on deep learning based brain computer interface: Recent advances and new frontiers. *J Neural Eng.* DOI: 10.1088/1741-2552/abc902. Epub ahead of print. PMID: 33171452.

[251] Xiang Zhang, Lina Yao, Dalin Zhang, Xianzhi Wang, Quan Z. Sheng, and Tao Gu. 2017. Multi-person brain activity recognition via comprehensive EEG signal analysis. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.* ACM, 28–37.

[252] Xiao-Lei Zhang and Ji Wu. 2012. Deep belief networks based voice activity detection. *IEEE Trans. Aud. Speech Lang. Process.* 21, 4 (2012), 697–710.

[253] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2017. Energy-based generative adversarial network. In *the 5th International Conference on Learning Representations (ICLR'17).*

[254] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. 2001. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* 16, 6 (2001), 582–589.

[255] Weilong Zheng, Jiayi Zhu, and Baoliang Lu. 2018. Identifying Stable Patterns over Time for Emotion Recognition from EEG. *IEEE Trans. Affect. Comput.* (2018). https://doi.org/10.1109/TAFFC.2017.2712143

[256] Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. 2014. EEG-based emotion classification using deep belief networks. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'14).* IEEE, 1–6.

[257] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010. Understanding transportation modes based on GPS data for web applications. *ACM Trans. Web* 4, 1 (2010), 1.

[258] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th International Conference on Ubiquitous Computing.* 312–321.

[259] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. *Proceedings of the ACM international Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'08).* 312. https://doi.org/10.1145/1409635.1409677

[260] Baoding Zhou, Jun Yang, and Qingquan Li. 2019. Smartphone-Based Activity Recognition for Indoor Localization Using a Convolutional Neural Network. *Sensors* 19, 3 (2019), 621.

[261] Pengfei Zhou, Yuanqing Zheng, Zhenjiang Li, Mo Li, and Guobin Shen. 2012. IODetector: A Generic Service for Indoor Outdoor Detection. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys'12).* ACM, 113–126. https://doi.org/10.1145/2426656.2426668

[262] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision.* 2223–2232.

[263] Xiaodan Zhu, Parinaz Sobihani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In *Proceedings of the International Conference on Machine Learning.* 1604–1612.

[264] Han Zou, Yuxun Zhou, Jianfei Yang, Hao Jiang, Lihua Xie, and Costas J. Spanos. 2018. Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network. In *Proceedings of the IEEE International Conference on Communications (ICC'18).* IEEE, 1–6.

Received June 2020; revised June 2021; accepted June 2021