Kun Qian, Zixing Zhang, Yoshiharu Yamamoto,
and Björn W. Schuller

Artificial Intelligence Internet of Things for the Elderly

From assisted living to health-care monitoring

An aging population is increasingly prevalent in both developed and developing countries, raising a series of social challenges and economic burdens. In particular, more elderly people are staying alone at home than are living with people who can take care of them. Therefore, assisted living (AL) and health-care monitoring (HM) can be critical issues in this era of human-centered artificial intelligence (AI). In this context, we aim to provide an encompassing review summarizing the state-of-the-art works combining AI and the Internet of Things (IoT) to help the elderly live easier and better. We systematically and comprehensively compare paradigms in terms of methodologies and application scenarios. The pros and cons among these technologies are discussed in detail. Then, we summarize current achievements and indicate their limitations. Finally, perspectives on highly promising future work are presented.

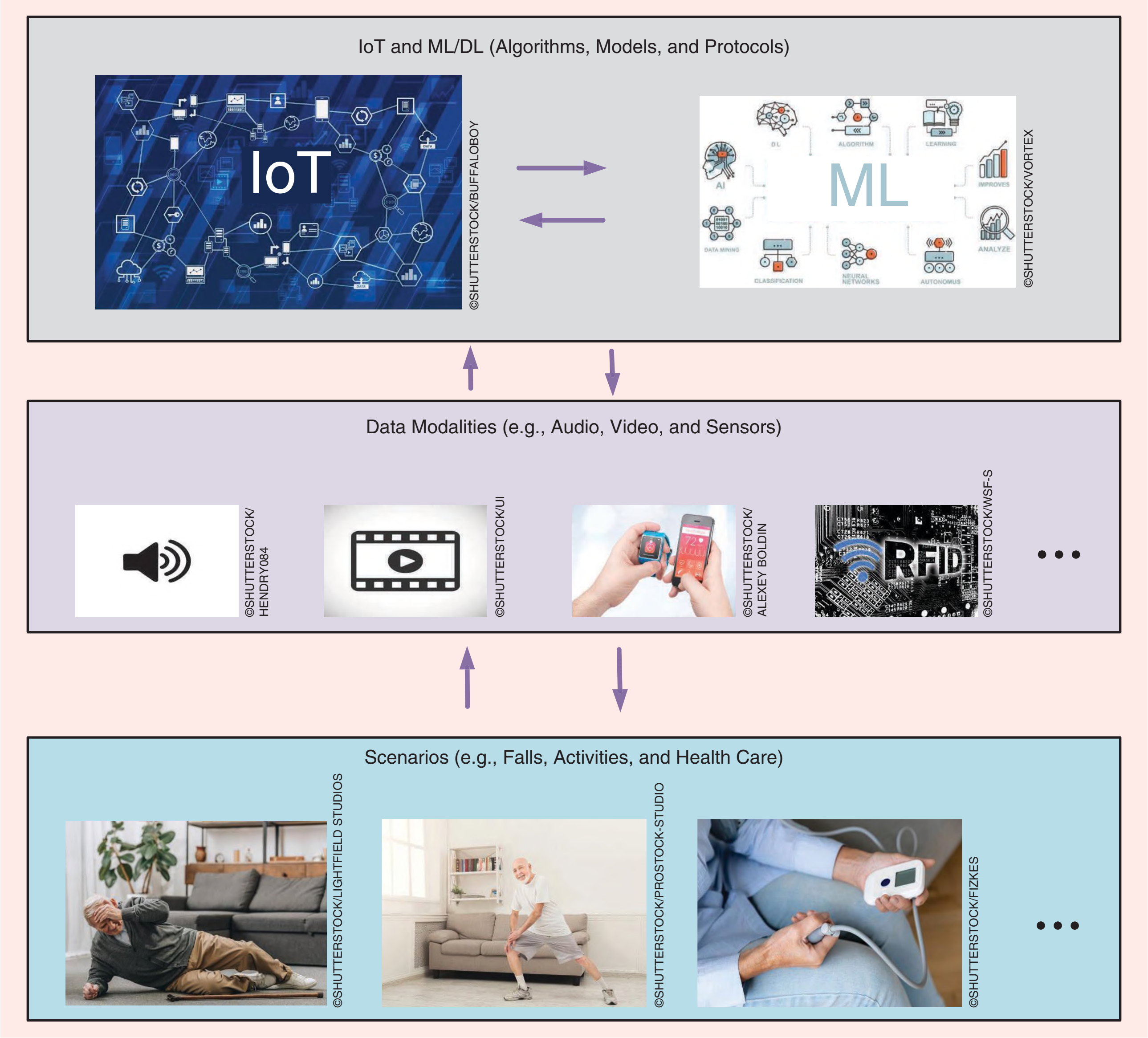
Overview

According to a report [1] by the World Health Organization, an aging population has become more and more prevalent in both developed and developing countries. Taking Japan as an example, approximately 27.6% of the citizens are already 65 or older [2], which makes the whole society face a series of economic burdens and social challenges. In this era of human-centered AI (HAI), we have witnessed tremendous efforts in the fields of AL and health monitoring that have been made by leveraging the power of AI and the IoT, which can be referred to by the acronym *AIoT* by combining the two crucial factors in the fourth Industrial Revolution. We can see promising achievements using ubiquitous sensors combined with state-of-the-art signal processing (SP) and machine learning (ML) techniques. Together, they can facilitate an easier, higher-quality life for individuals who suffer from chronic diseases and need special assistance. In particular, the elderly constitute a large market in our societies. To the assist the global fight against COVID-19, the AIoT can contribute solutions for eldercare by fully mining smart home sensor data.

An overview of AIoT studies related to eldercare can be found in Figure 1. The fundamental motivation should come from scenarios in the daily life of the elderly, e.g., falls and activity recognition. Then, the data modality that can be best used for a specific scenario will be considered. Finally, the IoT and ML/deep learning (DL) can enable intelligent systems. When reading the literature about AIoT applications for the elderly, we may find that there are two main directions, i.e., AL and HM. Analyzing the elderly's daily activities and providing proactive care are essential factors for applications in AL, while monitoring the health status and predicting the future health conditions of the elderly who are suffering from chronic diseases are the key challenges for HM studies.

> **An aging population is increasingly prevalent in both developed and developing countries, raising a series of social challenges and economic burdens.**

On the one hand, we have witnessed the great success achieved in the information and communications technology (ICT) field, e.g., 5G, the IoT, AI, big data, and cloud computing. On the other hand, the AIoT for the elderly is still a young field and underestimated. In particular, considering the specific demands of older individuals, we are still far from building a smart society and smart homes for our parents and even for ourselves. For instance, fall detection (FD) is an essential topic in AL care for the elderly who live independently. A plethora of efforts has been made to discover efficient modalities and methods. Yet, the state of play remains limited to presenting a robust paradigm that can be applied in the real world. Moreover, daily life behavior analysis is believed to be useful in activity recognition for


IoT and ML/DL (Algorithms, Models, and Protocols)
IoT
ML
DL
ALGORITHM
LEARNING
AI
IMPROVES
DATA MINING
ANALYZE
CLASSIFICATION
NEURAL NETWORKS
AUTONOMUS
Data Modalities (e.g., Audio, Video, and Sensors)
RFID
• • •
Scenarios (e.g., Falls, Activities, and Health Care)
• • •

©SHUTTERSTOCK/BUFFALOBOY

©SHUTTERSTOCK/VORTEX

©SHUTTERSTOCK/HENDRY084

©SHUTTERSTOCK/UI

©SHUTTERSTOCK/ALEXEY BOLDIN

©SHUTTERSTOCK/WSF-S

©SHUTTERSTOCK/LIGHTFIELD STUDIOS

©SHUTTERSTOCK/PROSTOCK-STUDIO

©SHUTTERSTOCK/FIZKES

FIGURE 1. An overview of the AIoT for the elderly. DL: deep learning; RFID: radio-frequency identification.

applications such as feasible alarm systems for emergency treatment. Nevertheless, how to protect personal privacy and maintain a long-term, low-energy, ubiquitous system is challenging. Last but not least, much more attention has been given to physical rather than mental health care for the elderly. In fact, we cannot fully determine the elderly's potential psychiatric issues.

While there are some previous surveys of ICT applications for the elderly, they fail to focus on a comprehensive investigation of SP and ML techniques in their corresponding fields, nor do they provide insights into and perspectives of the state-of-the-art works in the AIoT for dealing with the aging population challenge. For example, a recent survey introduced dense sensing network-based anomaly detection [1], presenting a detailed description of technologies used in home-based eldercare. However, this topic is only a subfield of general AIoT applications for the elderly, which limits the readership and the larger picture. Another work focused on smart homes for aging in place [2], giving a good summary of diverse application drivers but lacking an in-depth analysis of SP and ML techniques for mining sensor data. Unlike previous surveys, we focus on a comprehensive investigation of SP and ML methods applied to the AIoT for the elderly in terms of methodologies and scenarios. We also aim to provide a clear picture of state-of-the-art works and their limitations. As guidance and a tutorial, perspectives of future work will be included. Figure 2 describes the article content.

Assisted living and health-care monitoring can be critical issues in this era of human-centered artificial intelligence.

Modalities and scenarios

In this section, we illustrate the main data modalities and their applied scenarios for the elderly.

Data modalities

A variety of data modalities has been used for AL and HM applications for the elderly. Generally speaking, the data can be categorized as wearable and nonwearable. In this study, we summarize the main modalities used in the AIoT for the elderly according to their signal characteristics, which include, but should not be limited to, the following.

Audio

Audio data and related computer audition (CA) technology inherently have a noninvasive and ubiquitous character, so they can be widely used for in- and out-of-home surveillance-based applications in the eldercare field. As an example, Li et al. proposed an acoustic system that can automatically detect falls and promptly report accidents to caregivers [3]. In the study, a circular microphone array was used to capture the spatial information of sound signals, which can be employed to locate the near-field source signal via a steered response power technique and enhance the signal through a suitable beamforming (BF) method. Furthermore, a simple ML model, namely the k-nearest neighbor (k-NN), where $k = 1$, was used to classify the signals by extracting the mel scale frequency cepstral coefficients (MFCCs). Apart from passive sound data, speech can be regarded as an important information carrier to reflect the physical and mental health status of the elderly. In the Conference of the International Speech Communication Association Computational Paralinguistics Challenge (COMPARE) series, speech is, for example, used to estimate the neurological state of Parkinson's patients (in a regression task) [4] and classify the emotion of elderly people (in a classification task) [5], respectively. For these tasks, classic ML models (needing handcrafted features) and DL methods (owning high-level representations extracted from DL models) can be options.

Video

Computer vision (CV) and its related technologies occupy an important position in the AIoT for eldercare applications. Yu et al. introduced a CV-based method for FD by analyzing postures recorded by a single camera [6]. The authors included a codebook background subtraction algorithm to improve the results. For classification, ellipse fitting and a projection histogram were used to form the feature vectors, and a directed acyclic graph (DAG) support vector machine (SVM) was adopted as the classifier in a multiple classification scheme (bending, lying, sitting, and standing). Alaoui et al. proposed an algorithm to detect falls by extracting spatial and temporal features from videos [7]. In this method, the key points and skeletons of the human body were first detected. Then, the distances and angles between two pairs of sequential points were calculated. Principal component analysis was used to unify the feature dimensions. Finally, the authors compared four classic ML models, i.e., an SVM, the k-NN, a decision tree (DT), and the random forest (RF), in which the SVM outperformed the other three.

Another 3D skeleton-based approach for describing the spatial and temporal aspects of a human activity sequence was introduced in [8]. The Minkowski and cosine distances between 3D joints were used to represent the spatial variation

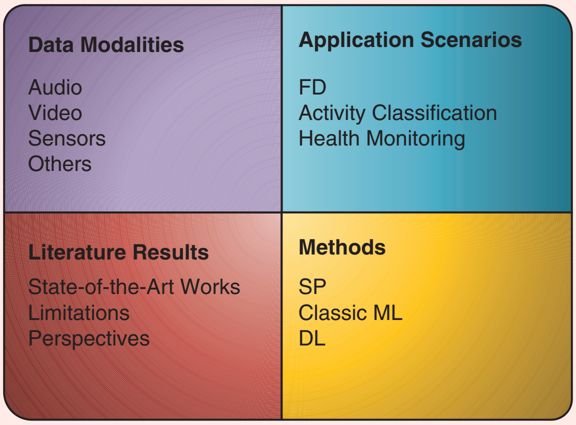


FIGURE 2. The content of this article.

of the activity sequence. Meanwhile, the difference in coordinates for the frame sequence between each 3D joint and both the maximum and the minimum values of the same 3D joint across the entire sequence were calculated to represent temporal variation in the activity sequence. An extremely randomized trees (ERT) algorithm was employed for the classification step. Chen et al. introduced an activity encoding scheme by which a skeleton sequence could be estimated from red–green–blue (RGB) images [9]. In the study, the skeleton sequence was generated from RGB images via a real-time pose estimation algorithm. Then, an interframe matching algorithm (IMA) was employed to filter the nontarget objects. Data augmentation was involved to enrich the small-scale skeleton sequences. Subsequently, activity images were generated from the gray matrix encoded from the skeleton sequences. Finally, a convolutional neural network (CNN) was used to fulfill the classification task.

To the assist the global fight against COVID-19, the AIoT can contribute solutions for eldercare by fully mining smart home sensor data.

Sensors

The fast development of sensors [3] and wearables makes it feasible to collect real-world information in terms of temperature, humidity, illumination, vibration, and human vital signs (heart rate, blood pressure, skin conductance, temperature, movement, gait, and so on). Furthermore, by leveraging the power of AI, sensor data can play an important role in the AIoT eldercare applications market. Khandoker et al. recorded foot clearance data when participants walked in a steady state on a treadmill [10]. A predefined minimum foot clearance (MFC), i.e., the value of the minimum vertical distance between the lowest point under the front part of the shoe or foot and the ground, was used as the gait variable. Wavelet-based features were demonstrated to be superior to statistical ones (both extracted from the MFC) when training an SVM. In addition, the models can also be used to estimate the relative risk of falls when calculating the calibrated posterior probabilities of the SVM. A comprehensive investigation of fall risk prediction capabilities using two types of wearable sensors (accelerometers and pressure-sensing insoles) was given in [11]. The authors also studied four accelerometer locations: the head, pelvis, and left and right shanks. In addition, they compared three types of ML models: a neural network (NN), an SVM, and a naive Bayes (NB) classifier. The NN model using dual-task gait data and features extracted from head, pelvis, and left shank accelerometers yielded the best performance.

Gochoo et al. used data collected from passive infrared (PIR) sensors to analyze the elderly's travel patterns according to a Martino–Saltzman model in a nonprivacy-invasive scenario [12]. The authors compared their proposed deep CNN (DCNN) with NB, SVM, *k*-NN, DT, gradient boost, RF, and one-versus-rest models. The DCNN and the RF were found to be the best for inferring dementia through travel pattern matching. In addition, the PIR motion sensors and door sensors were found to be efficient for monitoring the activities of a single elderly woman in a smart home for eight months [13]. In the study, the DCNN showed a capacity to extract intrasensor patterns from activity images converted from recorded PIR and door sensor data. To detect gait anomalies among patients suffering from Parkinson's disease, a deep time series-based approach was introduced in [14]. A hybrid architecture with a deep NN, including CNN layers as the reduction layer and a recurrent NN (RNN) with long short-term memory (LSTM) cells, plus a multilayer perceptron (MLP) as the classification layer, was used to analyze the acceleration values of the elderly's movements. Those accelerometers were placed in subjects' shoes, and the temporal time series were recorded as the sensor data.

A wrist-type three-axis accelerometer (i.e., a wristband) was used in [15]. In the study, a two-stage method combining RF classification and activity similarity calibration was found to be efficient for improving the recognition of elderly people's activity at home. Aoki et al. used a Kinect sensor to capture the elderly's whole-body movements, which were separated into a time series of 3D coordinates of body joints [16]. They used Hilbert–Huang transformation (HHT) and an SVM to analyze gait features recorded from the aforementioned data, which can be a good predictor of cognitive impairment among elderly individuals. Inertial sensors were used in [17], where four sensor locations (chest, lower back, wrist, and thigh) were explored (with an SVM) to classify the elderly's physical activities. The authors found that the sensor worn on the lower back achieved the best performance among the single-sensor solutions. By adding another sensor on the thigh, further improvement could be achieved, although including more sensors yielded no better results. Alkhatib et al. studied the feasibility of using kinetic data, e.g., the vertical ground reaction forces collected from various sensors underneath the foot, to analyze the elderly's gait [18]. It was demonstrated that spatial and time signal analysis methods combined with a linear discriminant analysis (LDA) classifier can be useful for detecting the balanced gait of subjects suffering from Parkinson's disease. Yu et al. found that to build a personalized health monitoring system for elderly people, smart wearable sensors can be necessary to transmit vital signs and physiological changes [19].

As indicated in [20], radar and related technologies are important for FD and health monitoring in AL, due to a series of inherited characteristics, e.g., proven technology, privacy preservation, nonintrusive sensing, nonobstructive illumination, insensitivity to lighting conditions, and safety. Radar frequency changes generated by the backscatters from people in motion, also known as *Doppler effects*, can carry prominent features that reveal different human motions and gross motor activities [20]. Su et al. proposed a Doppler range control radar that aims to detect falls among elderly residents [21]. Wavelet transformation (WT) and the *k*-NN ($k = 1$ in the study) were used. The general paradigm contained two stages: first, the WT coefficients at a given scale were used to identify the possibility of a fall; second, the WT coefficients at several scales across many successive frames were employed to form the feature vectors for the classification of fall and nonfall activities. Shrestha et al. introduced another S-band radar system for classifying the elderly's activities in daily life [22]. The authors used the traditional fast Fourier transform (FFT) to generate spectrograms from raw data, by which a series of time–frequency features can be extracted (and selected) for training an SVM model. They found it necessary to select frequency bands close to human movements.

Others

Apart from the aforementioned modalities, we can see other successful applications in the AIoT for the elderly's use. Personal sociodemographic and health-related factors were used as features for building an ML (the RF was selected) model for predicting anxiety and depression in elderly patients [23]. It provided to be a feasible method to implement a long-term mental healthcare management system by using data (age, gender, chronic medical conditions, and so forth) that can be easily collected from ubiquitous devices (e.g., smartphones). Bertini et al. studied the use of routinely collected socioclinical data, e.g., vital statistics and health information, to predict frailty in the elderly [24]. The authors found that a logistic regression model achieves the best performance among other ML models, e.g., an SVM and the RF.

> **We are still far from building a smart society and smart homes for our parents and even for ourselves.**

Application scenarios

In this section, we discuss application scenarios in the literature. To avoid duplicates, we focus on how these applications could help older people enjoy an active and independent life.

FD

Falls can result in injuries (causing personal suffering as well as the potential for a high economic cost), and they are a leading cause of death among the elderly [20]. Therefore, automatic FD is a crucial research direction in AIoT applications for the elderly. Table 1 summarizes and compares the main contributions from AIoT applications for FD. The current literature focuses on using classic ML models (e.g., an SVM, the *k*-NN, and the RF) and feedforward NN (FNN) models that have a shallow layer architecture. The reason for the lower network depths is that FD-related data are difficult to collect, which results in small data sets, hence restraining the capacity of DL models to learn sufficiently generalized representations from the inputs. In the paradigm of classic ML, specific human domain knowledge is a prerequisite for building a feasible model to fulfill a task. As one can see, advanced SP methodologies, e.g., WT (see Table 1), are an essential part of FD systems. When considering the data modalities, audio (CA)- and video (CV)-based methods appear quite efficient in the FD scenario. However, both may raise the issue of privacy intrusion, especially CV-based models. Wearable sensor- and radar-based methods appear better at protecting personal privacy and show promising results (see Table 1). Nevertheless, wearable sensors may have power management problems and inconvenience elderly people who have to carry them all day [3]. Furthermore, Doppler radar-based methods may misinterpret normal activities, such as a pet jumping and a person sitting down on a chair [20].

Activity classification

The automatic classification of daily activities is a crucial part of ambient AL technologies [25], enabling elderly people to live independently and facilitating the early detection of diseases, e.g., Alzheimer's and dementia. Tasks can be divided

Table 1. A comparison of studies of AIoT FD applications.

Reference	Modality	SP	ML/DL	Result	Findings
[3]	Audio	MFCC BF	1 NN	Sensitivity: 100% Specificity: 97%	BF can improve specificity; the proposed model is robust against different acoustic environments and floor materials.
[6]	Video	Codebook CV features	DAG SVM	DR: 97.1% FDR: 1%	The proposed method is robust to background noise; multiple moving objects and occlusions are challenges that needing to be addressed.
[7]	Video	CV features	SVM, DT RF *k*-NN	WAR: 98.5% Sensitivity: 97% Specificity: 100%	The best performance is achieved by the SVM; the method cannot be used in a dark room; the algorithm is time consuming.
[10]	Sensors	WT	SVM	WAR: 100%	Wavelet-based features achieved higher accuracy than the statistical features did; the models could also be used for fall prevention.
[11]	Sensors	Features	FNN NB SVM	WAR: 57% Sensitivity: 43% Specificity: 65%	The best model is achieved with dual-task gait data collected from head, pelvis, and left shank accelerometers; similar fall risk model performance can be reached using single- and dual-task gait assessments.
[21]	Sensors	WT	1 NN	AUC: 0.82–0.96 Sensitivity: 92.3–97.1% Specificity: 81.4–92.2% WAR: 83.5–93%	WT-based features are more robust than MFCC-based features for FD; WT features can also be used to identify the possible occurrence of a fall; the best performance will decrease in real-world settings (such as bathrooms and apartments) but are still acceptable.

Results are shown for the best approaches.
WAR: weighted average recall; UAR: unweighted average recall; DR: detection rate; FDR: false detection rate; AUC: area under the curve; FNN: feedforward NN.

into several categories, from low-level, easily recognized events (e.g., cooking, eating, and sleeping), to high-level abnormalities in physical and mental statuses. When looking at the data modalities used for activity classification (AC), sensors dominate (see Table 2). On the one hand, existing studies show encouraging results, as most models can reach more than 90% of the weighted average recall (WAR), also known as *accuracy*. On the other hand, one of the biggest challenges is the generalization of SP and ML methods across households, which still needs to be addressed [25]. For instance, when a visitor appears in a room, a previously trained model cannot be directly used [12].

Classic ML models can be efficient for a variety of AC tasks. At the same time, DL models have been increasingly studied in recent years and shown great potential to improve the performance of the current state of the art [9], [12]–[14]. Advanced SP technologies are quite useful for designing features in terms of classic ML methods (see Table 2). For DL models, a CNN is the most frequently selected architecture, due to its strong capacity to extract local and global features from data. It can be seen that adding an RNN structure to a hybrid DL architecture could be better because of the time series characteristics of the elderly's activities (see [14]). It is reasonable to think that capturing contextual information from activity data can improve the analysis of the elderly's behavior in daily life.

Health-care management

Considering the typical characteristics of the fast-growing elderly population [1], i.e., living alone (most elderly people prefer living in their own home even though they could have many options, such as nursing homes), cognitive impairments, chronic diseases, and vision and hearing constraints, it appears clear that HM plays a crucial role in the current and future AIoT application area. Relevant studies covering the AIoT for HM are summarized in Table 3. Compared to

Table 2. A comparison of studies of AIoT AC applications.

Reference	Modality	SP	ML/DL	Result	Findings
[8]	Video	CV features	ERT	WAR: 80.9, 73.4%	The device has a low cost, and the model can be used for real-time monitoring; the number of activities that the model can recognize is limited.
[9]	Video	IMA Activity encoding	CNN	WAR: 100% WAR: 100%	The activity encoding method is robust against an incomplete skeleton; during DL training, data augmentation is needed for small-scale skeleton sequences.
[12]	Sensors	Features	CNN	WAR: 97.8%	A few episodes with a large number of movements may reduce the accuracy of the classifier; the method cannot be used when a visitor is in a house.
[13]	Sensor	Features	CNN	WAR: 98.5% F1 score: 0.79	The model is bad at recognizing dish washing and meal preparation, due to the fact that those activities are similar in terms of locations and sensors.
[14]	Sensors	Features	CNN, MLP, LSTM–RNN	WAR: 95%	The RNN fully considers the characteristics of time series data; reducing the input dimensions can improve the results.
[15]	Sensors	Features ASM	RF	WAR: 95.6%	The activity similarity (the correlation between an activity, location, and time) should be considered and can improve the AC performance.
[16]	Sensors	HHT EMD	SVM	AUC: 0.77	The best performance is achieved when shoulder joint data are excluded; dual-task gait features are more effective than single-task features.
[17]	Sensors	Features	SVM	WAR: 96.8% F1-score: 0.88	The best performance can be achieved using two sensors' data (lower back and thigh); a feature selection step can reduce the computational cost.
[18]	Sensors	Correlation	LDA	WAR: 95%	The correlation is a simple but effective feature for gait analysis; the curvature radius of Parkinson's subjects is smaller than that of healthy subjects.
[22]	Sensors	FFT	SVM	WAR: 90%	Matching frequency bands to human movements during feature extraction is important for classifying specific events, e.g., falls.

Results are shown for the best approaches.
ASM: activity similarity matrix; EMD: empirical mode decomposition.

Table 3. A comparison of studies of AIoT HM applications.

Reference	Modality	SP	ML/DL	Result	Findings
[4]	Audio	CA features	SVR	ρ: 0.39	The severe acoustic mismatch appears because of the different recordings between partitions.
[5]	Audio	CA features	SVM CNN Transformer	UAR: 49.7%	The "bad confusions," i.e., low and high, are infrequent; V can be modeled better than A via linguistic features; a late fusion cannot yield the best performance.
[19]	Sensors	Features	LR, FNN SVM, DT	WAR: 68.1% Precision: 81.7%	The best performance is achieved by DT; the classification rules could be quite useful for health-care management.
[23]	Text	Features	RF	WAR: 91% Precision: 89%	Sociodemographic and medical factors can be used for predicting anxiety and depression via ML; a larger data set is needed.
[24]	Text	Features	LR	AUC: 0.7	The potential risk of missing data can be reduced through the use of routinely collected socioclinical data.

Results are shown for the best approaches.
ρ: Spearman's correlation coefficient; V: valence; A: arousal; LR: logistic regression; SVR: support vector regression.

the previous two application scenarios, the results in terms of the robustness and reliability of HM applications are modest. This can be explained by the inherited complicated characteristic of the topic itself. An efficient HM system needs to be built on massive data and long-term analysis, which requires more powerful SP and ML technologies, compared to FD and AC tasks. CA-based methods have shown potential for the early detection of Parkinson's disease [4] and the analysis of the elderly's emotions [5]. In this scenario, speech should be regarded as a kind of physiological signal that carries information about a subject's physical and mental health. Sensors (other than audiovisual ones) also form an important component of HM management [19], and other modalities (e.g., text) show promising results in mental health care [23] and frailty prediction [24].

Paradigms and algorithms

We now introduce the paradigms of classic ML and DL. Furthermore, the algorithms are briefly discussed in terms of their specific tasks and applications.

Computer vision and its related technologies occupy an important position in the AIoT for eldercare applications.

Classic ML

In the paradigm of classic ML (see Figure 3), extracting handcrafted features is the first, crucial step after preprocessing (which refers to the front end). By considering both the data modality and the application scenario, a variety of SP technologies can be employed for a specific task. Fourier transformation (FT) and its variant, i.e., the short-time FT (STFT), have been demonstrated to be efficient in capturing time–frequency information from data. Take CA-based methods as an example. Their frequently used features, i.e., MFCCs [3], and more sophisticated large-scale feature sets (COMPARE [4], [5]), are based on the STFT approach. Moreover, for radar signal analysis, the STFT is a powerful tool [22]. The assumption of the STFT is that a longer-duration nonstationary signal can be divided into a series of stationary signals with shorter lengths (segments), which can be suitable for separately conducting FT in each segment to reveal frequency domain information. Then, the changing spectra of the whole signal through time can be represented by a spectrogram. Finally, predefined features (which need human expert domain knowledge) can be extracted from the STFT consequences of the analyzed signals, which are used for building the ML model (in a regression or a classification case). As a classic SP method, STFT-based features have shown robustness and efficiency in a plethora of AIoT applications for the elderly (see Tables 1–3).

However, the drawbacks and limitations of the STFT are obvious. We cannot perfectly optimize the time–frequency resolution tradeoff caused by the Heisenberg effect. To reach a better resolution in the frequency domain, we should use a longer window length (e.g., a Hamming window) to divide the whole signal, although the resolution in the time domain will be worse. To this end, more advanced SP methods are worth exploring. WT, as a multiresolution analysis tool, can help to reach a higher frequency resolution in the lower-spectrum band and a higher time resolution in the higher-spectrum band. In particular, for the FD task, a Doppler radar captures the entirety of a fall, and its output has similar dynamic characteristics—a short duration of high frequencies and a long period of low frequencies—which makes WT more suitable than the STFT for analyzing the signals [21].

Nevertheless, finding a suitable and efficient wavelet function is not an easy job since it relies on empirical experiments and specific data. The HHT provides another option: the signal is decomposed into a number of intrinsic mode functions via the empirical mode decomposition (EMD) method and then applied to the Hilbert spectral analysis. The HHT can preserve the characteristics of the varying frequency, which makes it suitable for analyzing nonstationary and nonlinear time series data. For analyzing the movements of body joints in gait analysis of elderly people, the HHT was found to be superior to WT for providing a sharper frequency resolution [16]. But the HHT faces challenges in addressing some inherent issues, such as the end effects and mode mixing.

Another essential aspect of real-world application is the aforementioned preprocessing of signals. We need to take the

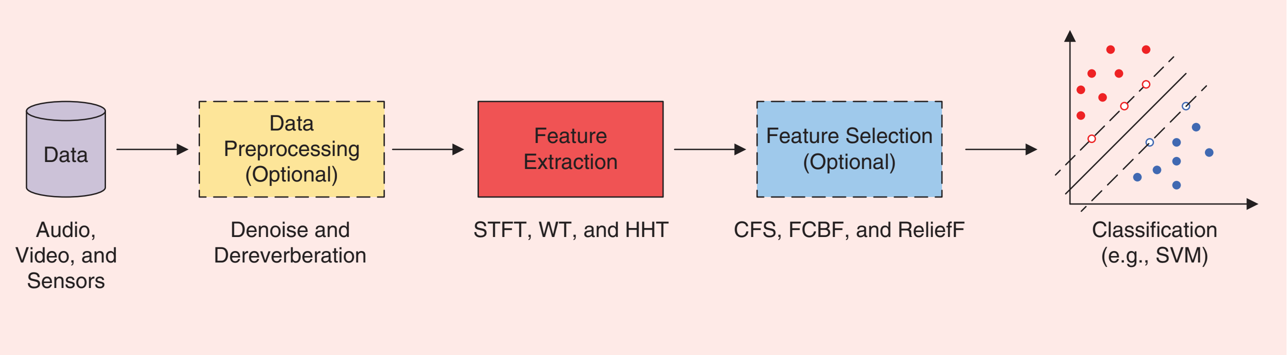

Data
Audio, Video, and Sensors
Data Preprocessing (Optional)
Denoise and Dereverberation
Feature Extraction
STFT, WT, and HHT
Feature Selection (Optional)
CFS, FCBF, and ReliefF
Classification (e.g., SVM)

FIGURE 3. The classic ML paradigm (using a classification task as an example). Feature extraction is a prerequisite for further model training and testing. Feature selection options can include correlation-based feature selection (CFS), a fast correlation-based filter (FCBF), ReliefF, and wrapper-based options [17]. STFT: short-time Fourier transformation.

complex environmental condition into account when designing an AIoT system for the elderly. Array SP technologies, e.g., BF, cannot be used only to reduce background noise; they must also attenuate any interference from directions other than that of the intended source signal [3]. A background subtraction algorithm is thought to be important in posture analysis based on CV methods [6]. Additionally, SP plays an important role in calculating specific features from data by using mathematical, physical, and statistical knowledge—denoted as *features* in Tables 1–3. Interested readers are referred to the "References" section.

When looking into the back end of classic ML, the SVM is the most popular model in the AIoT field. For its clear mathematical fundamental and stable performance across various data sets and tasks, the SVM has repeatedly been chosen as a standard baseline classifier [4], [5]. Training an SVM model entails finding the best hyperplane, or margin, that maximizes the separation between classes. An SVM can use a few samples (support vectors) from a training set to build the decision boundary that makes predictions. By adopting kernel functions, an SVM can be employed for analyzing linearly and nonlinearly separable problems. However, an SVM cannot perform very well when the data size is large and when the data set contains more noise (e.g., the target classes overlap). Besides, SVMs were originally designed for binary classification problems. This factor may restrain their capacity for multiclass problems, which are common in the AIoT for the elderly, although strategies can be adopted to make them feasible for recognizing more than two categories, e.g., "one versus one" and "one versus all."

Other simple classic ML models, such as the NB, LDA, logistic regression, the DT, and the *k*-NN, have been found to be efficient in several studies if the features are well designed and extracted. As one of several ensemble algorithms, an RF uses individual DTs to make a final prediction via a "bagging" algorithm. In this method, weak learners are first trained with randomly selected subsampled training data sets with replacement. Then, the final prediction will be made by a majority vote of the trained individual DTs. Therefore, the RF is thought to be more robust than using only a single DT model in real applications. An FNN as a shallow architecture among NNs can perform quite well if the data size is not large. Generally speaking, all the aforementioned classic ML models can be sufficiently powerful when handling small data sets, whereas they cannot significantly improve their performance with the large amounts of data usually produced by today's ubiquitous IoT sensors.

DL

Benefitting from the fast development of computational power during the past decade, DL has been applied to tremendous real-world AI applications and achieved notable successes in building efficient and robust models. In essence, DL is a series of nonlinear transformations of inputs, which results in the automatic extraction of high-level representations of the data. Considering the topology categories of DL models, one mainly finds that MLPs, CNNs, and RNNs prevail in the field (see Figure 4). An MLP is a kind of simple architecture that connects neurons (nodes) feeding forward (an FNN) from one layer to the next. Each layer is fully connected to all nodes of the subsequent layer, which is also known as a *fully connected* (*FC*) layer, whereas there are no connections between nodes within the same layer and across multiple layers.

A CNN is a combination of a series of convolutional layers (with a set of convolutional filters, also known as *kernels*), pooling layers, FC layers, and normalization layers, which can automatically learn features from images, audio, video, and text. An RNN is an architecture that features connections between neurons that can form a directed graph along a temporal sequence, which can learn context information by incorporating the outputs of a previous time step as the additional inputs for the current one. Considering the information flow direction, RNNs can be divided into unidirectional types (they have only a forward chain) and bidirectional variants (they have forward and backward chains). The procedure for training DL models is similar, namely, iteratively updating layer parameters to minimize the loss function, which measures the difference between the target outputs and the actual outputs. It can be executed through a backpropagation (BP) algorithm, and for an RNN, it can be performed by a BP-through-time (BPTT) algorithm. To overcome the vanishing gradient problem caused by BPTT, an LSTM or a gated recurrent unit cell can be used.

As discussed, DL has been incorporated into the AIoT for AC and HM application scenarios but not for FD (due to the extremely limited data size). Particularly for AC, the CNN dominates the choice of DL models (see Table 2) for its

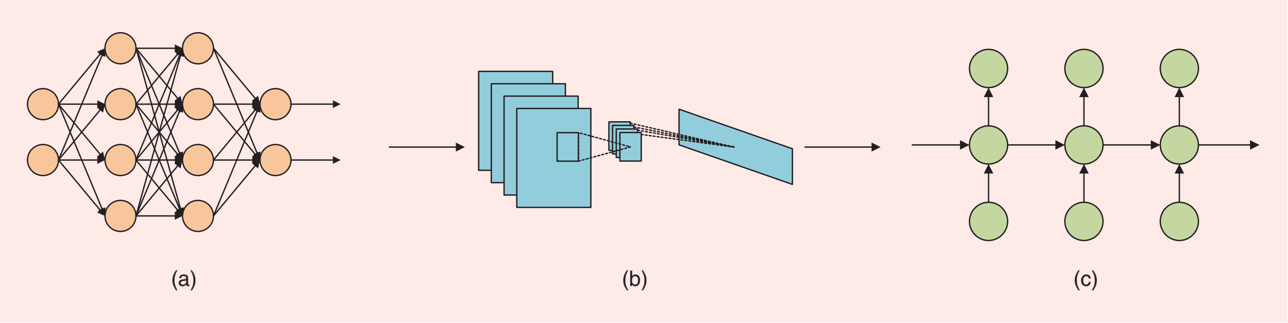


FIGURE 4. The main DL model topologies: the (a) MLP, (b) CNN, and (c) RNN.

strong capacity to extract high-level representations from data. However, we can also see that some hand-crafted features may still be needed when conducting these tasks. An RNN can improve model performance by capturing context information when analyzing time series data, such as in the cases of speech and gaits [14]. For the speech emotion recognition task in HM (see Table 3), we investigated and compared state-of-the-art DL models, such as deep transfer learning [5], sequence-to-sequence autoencoder-based unsupervised learning, and transformer-based linguistic feature learning. The best results were achieved by a deep transfer learning method for an arousal task (a three-class classification task referred to a scaled level of arousal) and the transformer-based method for a valence task (a three-class classification task referred to a scaled level of valence), respectively.

Where are we, and what is the future?

In this section, we first summarize the state-of-the-art works and their current achievements. Then, we analyze the limitations and challenges among those studies. Finally, we give our insights and perspectives for future directions.

State-of-the-art works

Most of the current studies have shown encouraging results even though the AIoT for eldercare applications is a young field. For the FD scenario, audio- and video-based methods can reach a very high FD sensitivity (better than 95%). In particular, by leveraging advanced SP technologies, e.g., BF and codebook background subtraction algorithms, the models can be made more robust to environments. WT can facilitate feature extraction from sensor and radar data, a process that is thought to be capable of providing multiresolution analysis of nonstationary signals (referring to signals related to falls). For the AC scenario, most of the studies have achieved a WAR better than 90% for recognizing a variety of elderly people's activities in daily life. Among the data modalities, sensor-based methods contribute the most to this area. More specifically, nonwearable sensors cannot well perform only in daily AC; they must also avoid most of the privacy intrusion issues caused by audio- and video-based methods. In addition, compared to wearable devices, they pose little inconvenience for elderly users. Computational human behavior analysis plays an important role in improving the performance of AC models. One can see that activity encoding, activity similarity, and context information are efficient factors that provide a higher-level representation than low-level descriptors in describing activities. For the HM scenario, both the classic ML and the state-of-the-art DL models have been investigated and compared. Handcrafted features may not be necessary if a DL model can automatically extract high-level features, whereas the robustness and generalization should be improved [5]. The management of chronic diseases, the early detection of dementia, and emergency care provision are the main directions in this domain.

Wearable sensors may have power management problems and inconvenience elderly people who have to carry them all day.

Limitations and perspectives

First, one should note that the existing studies were mainly conducted under ideal conditions, e.g., in labs and lab-like environments. Take FD task as an example. Most of the data were collected by stunt actors performing a fall in a lab, which could vary enormously from reality. Thus, to apply the methods in real-world situations, we need to consider in-the-wild data collection. Furthermore, some studies ignore subject independency in their experiments, which likely renders the trained models vulnerable to new data collected in a larger population size. Notice also that the WAR was widely used to evaluating models' performance in classification tasks. However, the WAR (or accuracy) is not a suitable evaluation metric for imbalanced data sets. In contrast, unweighted measures, such as the unweighted average, should be adopted when taking the imbalanced characteristic of a data set into account. In summary, the issues among the current studies likely cause overoptimistic results and expectations.

Another challenging issue is the lack of comprehensive investigations of data modalities that can be used for AIoT-based eldercare applications. Apart from the data modalities introduced in this article, it is worth exploring other kinds that may provide more opportunities in the field. For instance, the sense of touch (haptics) can be an essential sensory modality that facilitates daily activities [26]. We can imagine that by leveraging the power of AI (e.g., reinforcement learning), prostheses can be easier for the disabled elderly to use, resulting in faster and better rehabilitation. Furthermore, how to balance the power consumption and computational capacity of the hardware is another important issue. In real-world applications, AIoT designers should pay significant attention to system performance and costs.

Data scarcity is an inevitable challenge in this domain, particularly for the FD scenario. Data augmentation has shown a capacity to improve the generalization of DL models, although more advanced technologies should be involved. For instance, considering that labeled data are always limited, strategies such as unsupervised learning, semisupervised learning, active learning, reinforcement learning, and variants of the aforementioned algorithms should be explored. Human expert annotated data can still be expensive and rare even though the number of sensors is dramatically increasing. Moreover, generative adversarial networks can be used for both generating new samples that share the similar distribution of the original data set and for extracting more robust high-level representations of the data.

From the SP side, with the exception of advanced feature extraction methods, signal enhancement is also necessary in real-world applications. Take IoT-based microlocation in smart buildings as an example. Kalman filtering is an efficient method to overcome the interference effects of wireless devices [27]. Moreover, utilizing data collected from multiple sensors is another challenge. In a functional magnetic resonance imaging data analysis task, data fusion methods that enabled the exploitation of information shared across data sets were demonstrated to be superior to the data integration approach, which separately analyses data sets and combines the results [28]. Finding more suitable data fusion strategies is a promising direction, due to the increasing development of IoT sensors.

From the ML side, particularly for DL models, how to improve the robustness and generalization should be taken into account. A recent study demonstrated that using a noisy parallel hybrid DL model architecture can yield an accurate and robust estimation of remaining useful life [29]. Considering the socioclinical data mentioned in this article, more sophisticated DL architectures should be investigated. Another factor that cannot be ignored is the topology of DL models. It was shown that an encoder–decoder temporal convolutional network model can make more stable and accurate predictions than other framewise or sequential models in analyzing electromyographic signals [30]. Temporal information in time series signals carries important fingerprints of elderly people's behavior and daily life habits, which could be used for higher-level tasks, e.g., the early detection of disease and triggering emergency alarms. Moreover, fundamental studies of specific tasks are limited. For instance, human behavior analysis requires a combination of multiple disciplines, such as cognitive sciences, neuro/brain sciences, psychiatry, and AI, which cannot produce a solid conclusion for AC tasks. Particularly, for building an explainable and trustable AI system, we need more efforts from the broad scientific community.

Considering the ongoing COVID-19 epidemic and its effects on the elderly, AIoT-based applications can raise tremendous demands. First, the early detection of symptoms (coughing, pain, fevers, and so on) by IoT sensors combined with AI technologies can trigger an in-time warning and call for emergency care, particularly for elderly people who live alone. Second, monitoring the physical and mental health of the elderly can facilitate a secure social quarantine policy (e.g., 14 days at home). Last but not least, remote health center feedback obtained by analyzing big data via AI can maintain elderly people's confidence while they fight through a difficult period.

Conclusions

In this article, we gave a comprehensive overview of the AIoT applied to AL and HM for the elderly. Data modalities and application scenarios were presented in detail along with SP and ML algorithms. We highlighted the state of the art and indicated future directions in research fields. We hope this contribution can attract more attention and effort from academia and industry to work toward HAI applications to fight the population aging crisis.

Acknowledgments

This work was partially supported by the Japan Society for the Promotion of Science Postdoctoral Fellowship for Research in Japan (grant P19081); grants in aid for scientific research (19F19081, 20H00569, and 17H00878) from the Japan Ministry of Education, Culture, Sports, Science, and Technology; the Zhejiang Lab's International Talent Fund for Young Professionals (Project Hanami), China; and the European Union's Horizon 2020 Smart Environments for Person-Centered Sustainable Work and Well-Being project (grant 826506). Yoshiharu Yamamoto, and Björn W. Schuller share the senior coauthorship.

Authors

Kun Qian (qian@p.u-tokyo.ac.jp) received his Ph.D. degree in electrical engineering and information technology in 2018 from the Technical University of Munich. He is currently a Japan Society for the Promotion of Science postdoctoral research fellow in the Educational Physiology Laboratory, Graduate School of Education, University of Tokyo, Tokyo 113-8654, Japan. He serves as an associate editor of *IEEE Transactions on Affective Computing*, *Frontiers in Digital Health*, and *Bio Integration*, and he is a regular reviewer for numerous journals and conferences, including the International Conference on Acoustics, Speech, and Signal Processing and the Conference of the International Speech Communication Association. He has authored and coauthored more than 60 publications in peer-reviewed journals and conference proceedings, receiving more than 750 citations (h-index: 16). His research interests include signal processing, machine learning, biomedical engineering, and deep learning. He is a Senior Member of IEEE.

Zixing Zhang (connectzzx@gmail.com) received his Ph.D. degree in engineering from the Machine Intelligence and Signal Processing Group, Technical University of Munich, in 2015. He is with the Group on Language, Audio, and Music, Imperial College London, London SW7 2AZ, U.K. He has authored and coauthored more than 90 publications in peer-reviewed journals and conference proceedings, with more than 2,600 citations (h-index: 27). His research interests include semisupervised learning, active learning, and deep learning for applications in affective computing. He is a Member of IEEE.

Yoshiharu Yamamoto (yamamoto@p.u-tokyo.ac.jp) received his Ph.D. degree in education from the University of Tokyo in 1990. Since 2000, he has been a professor in the Graduate School of Education, University of Tokyo, Tokyo 113-8654, Japan, where he teaches and researches physiological bases of health sciences and education. He is an associate editor of *IEEE Transactions on Biomedical Engineering*, an editorial board member of *Technology* and *Biomedical Physics and Engineering Express*, and president of the Healthcare IoT Consortium. He has authored and coauthored more than 230 publications in peer-reviewed books, journals, and conference proceedings, leading to more than 11,000 citations (h-index: 57). His research interests include biomedical signal processing, nonlinear and statistical biodynamics, and health informatics. He is a Member of IEEE.

Björn W. Schuller (bjoern.schuller@imperial.ac.uk) received his Ph.D. degree in automatic speech and emotion recognition from the Technical University of Munich in 2006. He is a tenured full professor, heading the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and a professor of artificial intelligence, heading the Group on Language, Audio, and Music, Imperial College London, London SW7 2AZ, U.K. He is the field chief editor of *Frontiers in Digital Health*, former editor-in-chief of *IEEE Transactions on Affective Computing*, and president emeritus of the National Science Foundation Astronomy and Astrophysics Advisory Committee. He has authored and coauthored five books and more than 900 publications in peer-reviewed books, journals, and conference proceedings, leading to more than 35,000 citations (h-index: 86). He is a senior member of the Association for Computing Machinery and a Fellow of IEEE, the British Computer Society, and International Speech Communication Association.

References

[1] S. Deep, X. Zheng, C. Karmakar, D. Yu, L. Hamey, and J. Jin, “A survey on anomalous behavior detection for elderly care using dense-sensing networks,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 352–370, 2020. doi: 10.1109/COMST.2019.2948204.

[2] V. Nathan, S. Paul, T. Prioleau, L. Niu, B. J. Mortazavi, S. A. Cambone, A. Veeraraghavan, A. Sabharwal et al., “A survey on smart homes for aging in place: Toward solutions to the specific needs of the elderly,” *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 111–119, 2018. doi: 10.1109/MSP.2018.2846286.

[3] Y. Li, K. Ho, and M. Popescu, “A microphone array system for automatic fall detection,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1291–1301, 2012. doi: 10.1109/TBME.2012.2186449.

[4] B. W. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang et al., “The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson’s & eating condition,” in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 478–482.

[5] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos et al., “The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, breathing & masks,” in *Proc. INTERSPEECH*, Shanghai, P.R. China, 2020, pp. 2042–2046.

[6] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. Chambers, “A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1274–1286, 2012. doi: 10.1109/TITB.2012.2214786.

[7] A. Y. Alaoui, S. El Fkihi, and R. O. H. Thami, “Fall detection for elderly people using the variation of key points of human skeleton,” *IEEE Access*, vol. 7, pp. 154,786–154,795, Nov. 5, 2019. doi: 10.1109/ACCESS.2019.2946522.

[8] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal, “Skeleton-based human activity recognition for elderly monitoring systems,” *IET Comput. Vision*, vol. 12, no. 1, pp. 16–26, 2017. doi: 10.1049/iet-cvi.2017.0062.

[9] Y. Chen, L. Yu, K. Ota, and M. Dong, “Robust activity recognition for aging society,” *IEEE J. Biomed. Health Inf.*, vol. 22, no. 6, pp. 1754–1764, 2018. doi: 10.1109/JBHI.2018.2819182.

[10] A. H. Khandoker, D. T. Lai, R. K. Begg, and M. Palaniswami, “Wavelet-based feature extraction for support vector machines for screening balance impairments in the elderly,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 587–597, 2007. doi: 10.1109/TNSRE.2007.906961.

[11] J. Howcroft, J. Kofman, and E. D. Lemaire, “Prospective fall-risk prediction models for older adults based on wearable sensors,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1812–1820, 2017. doi: 10.1109/TNSRE.2017.2687100.

[12] M. Gochoo, T.-H. Tan, V. Velusamy, S.-H. Liu, D. Bayanduuren, and S.-C. Huang, “Device-free non-privacy invasive classification of elderly travel patterns in a smart house using PIR sensors and DCNN,” *IEEE Sensors J.*, vol. 18, no. 1, pp. 390–400, 2018.

[13] M. Gochoo, T.-H. Tan, S.-H. Liu, F.-R. Jean, F. S. Alnajjar, and S.-C. Huang, “Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN,” *IEEE J. Biomed. Health Inf.*, vol. 23, no. 2, pp. 693–702, 2019. doi: 10.1109/JBHI.2018.2833618.

[14] G. Paragliola and A. Coronato, “Gait anomaly detection of subjects with Parkinson’s disease using a deep time series-based approach,” *IEEE Access*, vol. 6, pp. 73,280–73,292, Dec. 19, 2018. doi: 10.1109/ACCESS.2018.2882245.

[15] H. Xu, Y. Pan, J. Li, L. Nie, and X. Xu, “Activity recognition method for home-based elderly care service based on random forest and activity similarity,” *IEEE Access*, vol. 7, pp. 16,217–16,225, Feb. 12, 2019. doi: 10.1109/ACCESS.2019.2894184.

[16] K. Aoki, T. T. Ngo, I. Mitsugami, F. Okura, M. Niwa, Y. Makihara, Y. Yagi, and H. Kazui, “Early detection of lower MMSE scores in elderly based on dual-task gait,” *IEEE Access*, vol. 7, pp. 40,085–40,094, 2019. doi: 10.1109/ACCESS.2019.2906908.

[17] M. Awais, L. Chiari, E. A. F. Ihlen, J. L. Helbostad, and L. Palmerini, “Physical activity classification for elderly people in free-living conditions,” *IEEE J. Biomed. Health Inf.*, vol. 23, no. 1, pp. 197–207, 2019. doi: 10.1109/JBHI.2018.2820179.

[18] R. Alkhatib, M. O. Diab, C. Christophe, and M. Elbadaoui, “Machine learning algorithm for gait analysis and classification on early detection of Parkinson,” *IEEE Sensors Lett.*, vol. 4, no. 6, pp. 1–4, 2020. doi: 10.1109/LSENS.2020.2994938.

[19] L. Yu, W. M. Chan, Y. Zhao, and K.-L. Tsui, “Personalized health monitoring system of elderly wellness at the community level in Hong Kong,” *IEEE Access*, vol. 6, pp. 35,558–35,567, July 19, 2018. doi: 10.1109/ACCESS.2018.2848936.

[20] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, “Radar signal processing for elderly fall detection: The future for in-home monitoring,” *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, 2016. doi: 10.1109/MSP.2015.2502784.

[21] B. Y. Su, K. Ho, M. J. Rantz, and M. Skubic, “Doppler radar fall activity detection using the wavelet transform,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 865–875, 2015. doi: 10.1109/TBME.2014.2367038.

[22] A. Shrestha, J. L. Kernec, F. Fioranelli, Y. Lin, Q. He, J. Lorandel, and O. Romain, “Elderly care: Activities of daily living classification with an S band radar,” *J. Eng.*, vol. 2019, no. 21, pp. 7601–7606, 2019. doi: 10.1049/joe.2019.0561.

[23] A. Sau and I. Bhakta, “Predicting anxiety and depression in elderly patients using machine learning technology,” *Healthcare Technol. Lett.*, vol. 4, no. 6, pp. 238–243, 2017. doi: 10.1049/htl.2016.0096.

[24] F. Bertini, G. Bergami, D. Montesi, G. Veronese, G. Marchesini, and P. Pandolfi, “Predicting frailty condition in elderly using multidimensional socioclinical databases,” *Proc. IEEE*, vol. 106, no. 4, pp. 723–737, 2018. doi: 10.1109/JPROC.2018.2791463.

[25] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, “Monitoring activities of daily living in smart homes: Understanding human behavior,” *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 81–94, 2016. doi: 10.1109/MSP.2015.2503881.

[26] T. Gathmann, S. F. Atashzar, P. S. Alva, and D. Farina, “Wearable dual-frequency vibrotactile system for restoring force and stiffness perception,” *IEEE Trans. Haptics*, vol. 13, no. 1, pp. 191–196, 2020. doi: 10.1109/TOH.2020.2969162.

[27] P. Spachos, I. Papapanagiotou, and K. N. Plataniotis, “Microlocation for smart buildings in the era of the internet of things: A survey of technologies, techniques, and approaches,” *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 140–152, 2018. doi: 10.1109/MSP.2018.2846804.

[28] Y. Levin-Schwartz, V. D. Calhoun, and T. Adali, “Quantifying the interaction and contribution of multiple datasets in fusion: Application to the detection of schizophrenia,” *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1385–1395, 2017. doi: 10.1109/TMI.2017.2678483.

[29] A. Al-Dulaimi, A. Asif, and A. Mohammadi, “Noisy parallel hybrid model of NBGRU and NCNN architectures for remaining useful life estimation,” *Quality Eng.*, vol. 32, no. 3, pp. 371–387, 2020. doi: 10.1080/08982112.2020.1754427.

[30] J. L. Betthauser, J. T. Krall, S. G. Bannowsky, G. Lévay, R. R. Kaliki, M. S. Fifer, and N. V. Thakor, “Stable responsive EMG sequence prediction and adaptive reinforcement with temporal convolutional networks,” *IEEE Trans. Biomed. Eng.*, vol. 67, no. 6, pp. 1707–1717, 2020. doi: 10.1109/TBME.2019.2943309.