IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 November 2022, accepted 23 December 2022, date of publication 5 January 2023, date of current version 10 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234421

RESEARCH ARTICLE

Video Based Mobility Monitoring of Elderly People Using Deep Learning Models

LAURA ROMEO, ROBERTO MARANI, TIZIANA D'ORAZIO, AND GRAZIA CICIRELLI
Institute of Intelligent Industrial Systems and Technologies for Advanced Manufacturing (STIIMA), National Research Council (CNR), 70125 Bari, Italy

Corresponding author: Grazia Cicirelli (grazia.cicirelli@stiima.cnr.it)

This work was supported in part by Programmi Operativi Regionali (POR) Puglia Fondo Europeo Sviluppo Regionale-Fondo Sociale Europeo (FESR-FSE) 2014–2020 through InnoNetwork Call titled "BESIDE: BEhavioral integrated System for diagnosis, support and monitoring of Neuro-Degenerative diseasEs" under Project YJTGRA7.

ABSTRACT In recent years, the number of older people living alone has increased rapidly. Innovative vision systems to remotely assess people's mobility can help healthy, active, and happy aging. In the related literature, the mobility assessment of older people is not yet widespread in clinical practice. In addition, the poor availability of data typically forces the analyses to binary classification, e.g. normal/anomalous behavior, instead of processing exhaustive medical protocols. In this paper, real videos of elderly people performing three mobility tests of a clinical protocol are automatically categorized, emulating the complex evaluation process of expert physiotherapists. Videos acquired using low-cost cameras are initially processed to obtain skeletal information. A proper data augmentation technique is then used to enlarge the dataset variability. Thus, significant features are extracted to generate a set of inputs in the form of time series. Four deep neural network architectures with feedback connections, even aided by a preliminary convolutional layer, are proposed to label the input features in discrete classes or to estimate a continuous mobility score as the result of a regression task. The best results are achieved by the proposed Conv-BiLSTM classifier, which achieves the best accuracy, ranging between 88.12% and 90%. Further comparisons with shallow learning classifiers still prove the superiority of the deep Conv-BiLSTM classifier in assessing people's mobility, since deep networks can evaluate the quality of test executions.

INDEX TERMS Deep neural networks, motion ability evaluation, skeleton based approach, video analysis.

I. INTRODUCTION

In recent years, the research on video analysis for human activity recognition has received an extensive boost for significant applications in various contexts, including surveillance, sports, human-machine interaction, rehabilitation, health monitoring, and robotics [1]. In particular, in the healthcare context, the analysis of human movements has allowed the realization of various functions such as remote diagnosis, support in the surveillance of fragile patients, recognition of anomalous events, etc. Many products and services have been developed for Ambient Assisted Living to aid healthy, active, and happy aging. The world is experiencing a rapid increase in the number of older people, which is expected to double over the next three decades [2]. Furthermore, there is an increasing spread of neurodegenerative diseases that heavily affect the well-being and healthy aging of the elder population [3]. As a consequence, elderlies need periodic monitoring to assess their movement skills. However, they are often unwilling to visit health clinics regularly, because of disabilities or logistical limitations, such as living in remote areas, thus wasting time, effort, and travel costs.

In this scenario, the analysis and control of people's motion and cognition abilities are fundamental in improving their social and clinical living conditions. Several studies demonstrate a strict link between cognitive impairment and motion dysfunction, including deficits in gait and balance [4], [5]. So, the study of human movements by video analysis can

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

significantly help assess people's motion abilities, providing objective evaluations and supporting remote diagnosis. Well-defined mobility tests exist in clinical contexts to assess people's mobility [6]. They consist of postural stability exercises, usually administrated and observed by physicians or specialized physiotherapists to measure people's functional mobility. Automatic video-based systems could greatly help to monitor these exercises in both home and clinical environments, obtaining objective and quantitative evaluations to support both expert personnel and medical diagnosis.

This paper proposes a vision-based system that observes elderly people while performing three well-defined mobility tests and automatically categorizes their mobility performance. In particular, the main contributions of this work are the following:

- The proposed system emulates the complex decision process of the expert physiotherapists in the evaluation of the mobility tests.
- The system processes real data acquired using low-cost commercial RGB cameras, typically implemented for video surveillance applications. The cameras were installed in two nursing homes that house older people who are healthy and affected by neurodegenerative diseases. The video data have been augmented and then processed to select the most informative features to provide a better-generalized model and enhance the decision process.
- Four classifiers with deep neural network architectures, based on Long-Short Term Memories (LSTMs) and Bidirectional Long-Short Term Memories (BiLSTMs), are proposed to classify the acquired data. The presented deep neural network architectures have also been rearranged to develop also regression models to further compare results with those from the classification task. Besides, comparisons with various traditional machine learning methodologies have also been conducted.

The remainder of this paper is structured as follows: Section II explores works more related to the proposed system; details about the case study are given in Section III; Section IV defines the different steps of the applied methodology; experimental results are in Section V, while final remarks are in Section VI.

II. RELATED WORKS

A. TECHNOLOGIES

In the existing literature, various instrumented systems have been proposed for real-time assessment of older people's mobility [4], [5], [7], [8]. The most relevant works have been summarized in Table 1.

Several works propose wearable sensors based on Inertial Measurement Units [9], or Inertial and Magnetic Measurement Systems for the evaluation of the physical functions of individuals [10], [11]. These sensors include accelerometers, gyroscopes, and magnetometers that measure the acceleration or angular velocity of the body segments to which they are attached. Although wearable sensors return valid information related to the movement of people, their output strictly depends on their position and orientation, and the activities to be monitored. Furthermore, older people and especially those suffering from neurological disorders do not easily accept unfamiliar devices.

Contrary to wearable sensors, non-wearable ones are non-invasive for people, as they are placed in the environment. Among the most commonly used for evaluating motion abilities, there are vision-based systems characterized by cameras that acquire video information of the human body and then, by using image processing techniques, extract relevant parameters useful for the analysis of motion abilities [12]. Marker-based Motion Capture Systems (MCSs), consisting of several cameras and a set of retro-reflective markers attached to the body of the monitored subjects, are an example of vision systems beneficial for capturing human movements with reliable accuracy [13]. However, high installation costs, expertise to set up and operate the system, and marker placement and calibration, limit their use in the home, and clinical environments [14]. Furthermore, the need for markers placed on the body brings out the same drawbacks of wearable systems. Typically, MCSs are used primarily in research laboratories or controlled environments to validate other sensory systems, such as webcams or RGB-D cameras, due to their high accuracy [15].

The limitations of marker-based systems have led to the development of markerless vision-based systems for human motion analysis [16]. In the last few years, the progress in new and low-cost optical technologies, together with the development of new and accurate pattern recognition approaches, has led to an increase in vision-based research works in this context [17], [18]. Monocular RGB cameras, stereo cameras, thermal cameras, and the recently developed RGB-D cameras, such as Microsoft Kinect or Intel RealSense [19], are the most commonly used systems to capture body movements and postural stability for assessing physical dysfunctions [20], [21].

A Kinect camera is used in [22] to observe older people while performing the Sit-to-Stand test to quantify the time taken to perform the test and to discriminate between elderly fallers and non-fallers in both laboratory and home assessments. A Kinect-based system has also been used to calculate the postural sway of older adults, estimating the variation of the center of mass of the body to provide a risk assessment of falls [23] or discriminate postural abnormalities [24].

A variety of vision systems have been used in the literature also for gait analysis. In [25], a Kinect camera and a neural network approach have been applied to identify the most significant gait characteristics and thus detect the disorders caused by Parkinson's disease. In [26], a low-cost thermal vision sensor has been proposed to continuously monitor the gait velocity of older people in their homes, showing a high correlation with that measured with a stopwatch. Two smartphone cameras have been used in [27] to classify normal

and abnormal gait. Different stages of Parkinson's disease, related to the severity of gait impairment, have been instead classified in [28] by using a system of two Kinect devices and by applying several classifiers, such as decision trees, Bayesian networks, neural networks, and K-nearest neighbor, for finding the most accurate one. In [29], a low-cost monocular RGB camera discriminates neurodegenerative patterns in the early stages of the disease.

B. METHODOLOGIES

In general, gathering data by observing people is not enough to assess the postural stability problem of human beings. Such information must be processed and elaborated through proper advanced systems to extract as much information as possible regarding the health of the elderly. In recent years, machine learning techniques for assessing movement skills are gaining more and more interest in the healthcare field [30], [31]. In particular, deep learning methodologies prove to be fundamental in health informatics. The development of automatic methods can lead to the generation, processing, and evaluation of complex data, which is difficult to deal with without the aid of technological systems. Table 2 gives an outline of the deep learning methodologies in the literature related to the presented work.

Several deep learning architectures have been used to process different types of data. Among them, the Convolutional Neural Networks (CNNs) are usually of significant impact in pattern recognition, from image to voice processing [32]. In [33], two types of CNN architecture, designed to analyze footprint pressure images from an instrumented walkway, have been compared to classify Huntington's disease severity. Similarly, a CNN was used in [34] to classify three severity stages of Alzheimer's disease using accelerometer data records. Considering the complexity of the classification problem and the presence of complex pattern sequences of mixed length, CNN seems suitable for managing this type of data and obtaining high accuracy rates for the three classes.

Alternatively, Recurrent Neural Networks (RNNs) are widely used for the analysis of time series in applications where the outputs depend on the previous computations, such as the analysis of text, speech, and movements. In [35], an RNN processes accelerometer signals to detect falls and estimate corresponding risks in real-time, reaching high efficiency and accuracy.

An evolution of the RNN is the Long Short-Term Memory (LSTM) network, which adds cell states to the network to expand the memory of the RNN [36]. In [37], the LSTM network has been applied to sequences of spatiotemporal gait parameters to capture both temporal variations and asymmetries in gait in patients with Parkinson's disease. LSTM network, taking advantage of remembering long-term dependencies within the data, achieves high accuracy rates [38].

In general, deep learning methods have several shortcomings. They typically have very complex architectures and time-intensive training phases. Furthermore, they need a large amount of data to reveal good performance. As a result, algebraic operations involving dense matrices, matrix products, and convolutions require equally enormous resources. Therefore, they must be transferred to Graphic Processing Units (GPUs) to accelerate machine learning processes [39]. However, compared to traditional methods, deep learning methods automatically learn hierarchical feature representations that capture their spatial and temporal correlations. In addition, such methods can approximate complex non-linear functions by composing several transformations of feature representations among the network layers from one level to more abstract levels.

C. OPEN CHALLENGES

From the literature analysis, it is evident that reported results are mostly simulated, whereas other works explore only partially the analysis of people's movements [40]. Several issues are still open:

- The use of video-based systems for classifying the motion capabilities of older people is not yet widespread in clinical practice. Many examples have been proposed to extract parameters for gait analysis, but no work has been found for clinical evaluations of motion skills. There is a need to develop vision-based motion analysis systems to collect accurate kinematic data in a non-invasive and valid manner to support the evaluation processes of medical staff in telehealthcare contexts.
- The poor availability of data does not allow the development and testing of classification methodologies to analyze human movements according to medical protocols. The literature shows many works that usually analyze human movements for binary classification, such as anomalous vs. normal behavior, healthy vs. sick people, or extract some parameters for disease recognition, disease stage classification, etc. The datasets in the literature are related to these limited classifications, while no data are available for the classification of motion abilities.

In light of these open challenges, there is a need for a non-invasive system to automatically assess the abilities of older people while performing motion tests of a clinical protocol. In this way, the intelligent system can emulate specialized clinical therapists who, observing the execution of the protocol, give a discrete score (class) to categorize the mobility level of older people.

III. CASE STUDY DESCRIPTION

The system setup used for data acquisition was made up of two low-cost RGB monocular cameras installed in two nursing institutes. One frontal camera and one side camera were placed in the gym of the institutes, where people usually execute mobility tests as shown in Figure 1.

The motion protocol, defined by medical staff and used in this work, consists of three mobility tests included in the so-called Short Physical Performance Battery (SPPB) [41]: the Balance Test (BT), the Walking Test (WT) and the Sit to

TABLE 1. Outline and limitations of the most relevant works concerning motion monitoring systems in the literature.

Reference	Technologies	Limitations
[9]	Survey that presents several wearable motion systems to assess gait in patients affected by Parkinson's disease.	Some of the presented works are only simulated. People suffering from neurological disorders do not easily accept unfamiliar devices.
[13]	The authors studied the gait patterns of patients affected by Alzheimer's disease, performing a motion and a cognitive task.	The paper gives the statistical information about the performance of patients in executing the tasks. However, there are no insights regarding a possible predictive model that automatically assesses the cognitive and motion impairment of the patients.
[22]	A Kinect camera is used to monitor older patients performing the Sit-To-Stand exercise, aiming to discriminate between non-fallers and potentially fallers.	The showed results are limited to statistical analysis, leaving open issues about a possible way to prevent patients from falling.
[25]	Patients with Parkinson's disease have been monitored using a Kinect camera, and the most significant gait features have been identified using neural network models.	Even though the paper presents an interesting analysis performed considering both patients and healthy subjects, all the analyses are limited to a single exercise.
[26]	A home monitoring system is proposed, where the gait velocity of older people is analyzed by merging information between a low-cost thermal vision sensor, and a stopwatch.	The work is focused on the real-time tracking of older people while they are in their homes. The issue of preventing older people from falling is not deepened.
[27]	A system composed of two smartphone cameras is used to classify the gait of the subject, discriminating between normal and abnormal.	The dataset has been gathered in a simulated environment, and there is no information about the health conditions of the subjects. Furthermore, only a single exercise has been analyzed.
[28]	Two Kinects are implemented to gather information about the severity of gait impairment in patients affected by different stages of Parkinson's disease. Several classifiers are considered in work.	The paper is focused on a single task that the patients have to perform. Only the gait is analyzed, thus ignoring other motion skills that could have been included to deepen the study.
[29]	A low-cost RGB camera system is implemented to monitor patients and subjects. A gait analysis has been performed, to automatically recognize whether a subject is affected by dementia, using different classifiers.	The healthy subjects and the patients affected by neurodegenerative diseases have a completely different range of ages. Furthermore, the work is focused only on one exercise.

TABLE 2. Outline and limitations of the deep learning methodologies found in the literature related to the presented work.

Reference	Methodologies	Limitations
[33]	The Huntington's disease severity is classified by comparing two CNN architectures that analyze the footprints of an instrumented walkway.	There is no comparison with classic machine learning models. Only the gait is analyzed.
[34]	The authors developed a method that processes accelerometer data acquired from patients with Alzheimer's disease, while they performed a gait exercise. Such data are fed to a CNN model to classify the stages of the disease.	Only one learning model has been analyzed, and all the study is based on a single exercise.
[35]	Accelerometer signals are processed with an RNN to estimate the risk of falls in real time.	The accelerometer signals are obtained using a wearable system, which can be invasive and uncomfortable for elderly patients.
[37]	Deep learning methods, particularly LSTM models, are used for fall risk assessment. Spatiotemporal sequences representing gait parameters are extracted from inertial sensors.	Wearable sensors are often unrecommended for elderly subjects. Furthermore, the exercises performed by the patient are limited to the ones that release gait information.

Stand Test (STST). Figure 2, shows a representative scheme of these tests. Specifically, in the Balance Test, the person stands with the feet side-by-side, then in a semi-tandem position and then in a tandem position, trying to stay in each of the listed positions for ten seconds (Figure 2 a)). In the Sit-To-Stand test, the person sits down and stands up five times with the arms crossed on the chest (Figure 2 b)). In the Walking Test, the person walks a four-meter linear path, free of obstacles, and returns to the starting point (Figure 2 c)).

The SPPB is usually administered to people by a physiotherapist to evaluate their mobility level as it releases information regarding body posture, balance, strength, and stability. The physiotherapist evaluates the execution of each test, giving a score value between 0 and 3, representing the mobility class. The classes range from the bad one (0 value), when the person cannot execute the test, to the best one (3 value) when, instead, the person succeeded.

All the older people, and their families, where needed, gave their written informed consent to participate in this study. There were 20 people affected by neurodegenerative diseases in the early stages and 27 healthy people, all in the range of age of 60 to 95 years. The subjects were recorded while performing the tests included in the SPPB in two separate acquisition campaigns three months apart. Several difficulties emerged during the data acquisition phase as the sample of people who participated in the first acquisition campaign was reduced in the follow-up as some were no longer able to perform the SPPB tests independently.

Once the video sequences of RBG images were acquired, they were appropriately processed to extract bidimensional skeletal data that have been made publicly

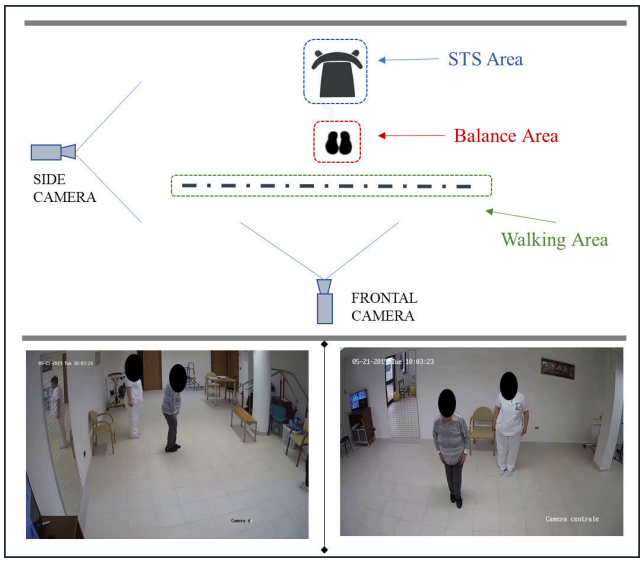

FIGURE 1. Camera setup used for video acquisition during the execution of the motion protocol.

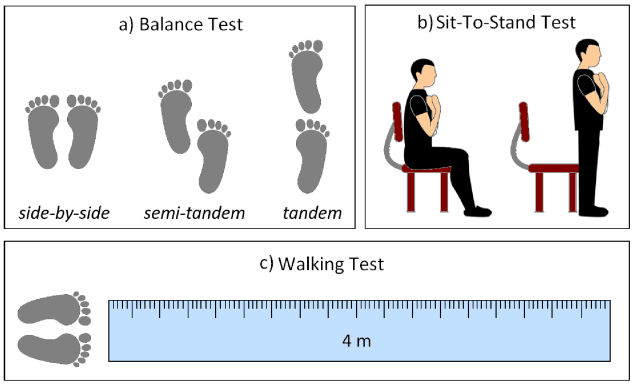

FIGURE 2. Representation of the three SPPB tests: a) Balance test: the patient stands with the feet side-by-side, then in semi-tandem and tandem positions; b) Sit-To-Stand Test: the patient sits down and stands up five times with the arms crossed on the chest; c) Walking Test: the patient walks for four meters.

TABLE 3. Statistical analysis of the videos of each test BT, WT, and STST, respectively.

Test	Total nr. of videos	Total nr. of Frames	Mean nr. of Frames	Standard Deviation
BT	74	19416	262.37	32.03
WT	76	10515	138.35	75.01
STST	96	39509	411.55	143.56

available [42], [43]. Table 3 gives some information about the acquired videos. In particular, 74, 76, and 96 videos have been captured for the BT, WT, and STST, respectively. As proved by the standard deviation values, the number of frames varies considerably among the three tests. For this reason, the duration of tests is not enough discriminant to achieve mobility assessment: a qualitative evaluation of test execution is mandatory to classify people's mobility.

IV. METHODOLOGY

The proposed system assesses people's mobility in the same classes defined by physiotherapists, but in a completely automatic and objective way, without human bias. The main steps involved in the proposed methodology are reported in Figure 3:

1) Commercial low-cost RGB cameras for video surveillance capture videos of test execution;
2) A preliminary processing extracts skeletal joints to evaluate complex details related to body postures, inclinations, and orientations of body parts;
3) A data augmentation technique enlarges the dataset made of the temporal evolution of joints in the image plane;
4) Significant features are extracted to construct input vectors to feed neural networks.

As primary output, this paper proposes deep neural network architectures for classification based on Long-Short Term Memory (LSTM) and Bidirectional Long-Short Term Memory (BiLSTM). Following an ablation study, preliminary convolutional blocks are added for feature mixing to improve classification results. Further comparison with standard classifiers from shallow learning (Decision Tree, Naive Bayes, SVM, KNN) and deep neural network architectures for regression, i.e. labeling people's mobility with continuous scores, are also presented. The next subsections will better detail the feature extraction process (Section IV-A), the network architectures used for classification (Section IV-B), and the data augmentation technique (Section IV-C).

A. FEATURE EXTRACTION

In this work, the well-known OpenPose library [44] is used to extract human skeletons from RGB frames. OpenPose efficiently detects the 2D pose of multiple people in an image, representing both the position and orientation of human limbs. The implemented model identifies 18 skeletal joints and 17 links between joints, as shown in Figure 4. Joint positions are not directly used to model people's mobility. Instead, a set of features is designed in agreement with clinicians to characterize anomalies during the SPPB tests. These features are based on 2D pairwise joint distances, normalized to body height, and geometrical angles between consecutive body segments (i.e. bones) to highlight posture variations and walking or balance problems. Features are evaluated at each frame and then put together in time series.

Figure 4 and Table 4 show the features of each SPPB test, providing detailed descriptions and indicating the camera used for their extraction. In the following, each SPPB test is analyzed together with the related features.

- *Balance/Walking Test:*
  Both balance and walking tests are administered to people to assess their static and dynamic skills. In the two cases, the following features have been considered:
  - Distance between feet (*(i)* in Figure 4). This distance can help evaluate the patient's confidence in following the predefined path of the WT.

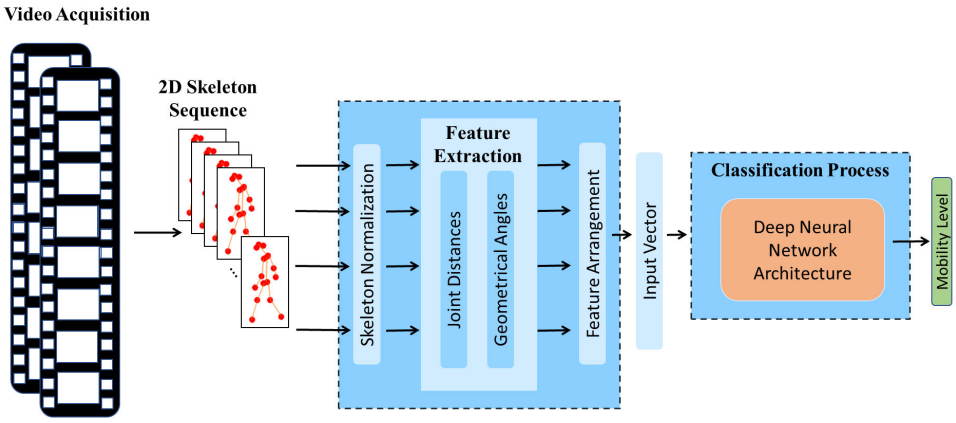

FIGURE 3. Pipeline of the proposed approach for classifying people's mobility level.

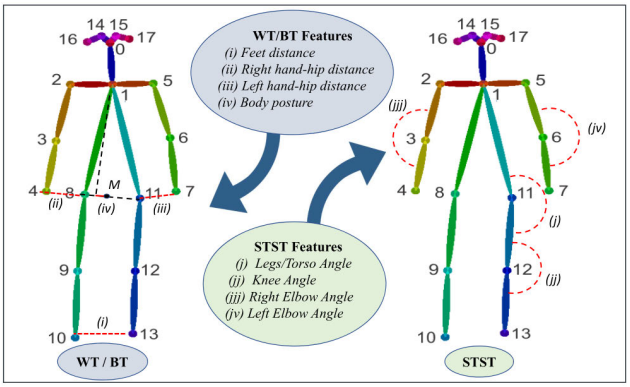

FIGURE 4. Features defined for the Walking and Balance Test (WT/BT) and Sit to Stand Test (STST), respectively.

- Distance between the right (or left) hand and the right (or left) hip from the frontal camera (*(ii)* or *(iii)* in Figure 4). This feature is fundamental for evaluating an eventual loss of balance and, in this case, for restoring balance with the help of the arms.
- Body posture, i.e. the column projection of the distance vector that connects the neck and the middle point M between the hips (*(iv)* in Figure 4). It provides information on people's torso inclination, indicating whether they keep their back straight.

It is worth noticing that, in the case of BT, the side-by-side, semi-tandem, and tandem tests are captured in three different videos. Homologous features are thus concatenated in vectors of increased lengths.

- *Sit-To-Stand Test:*
The STST is slightly different from the previous two tests, as it provides a method to quantify the functional strength of the lower limbs and/or to identify how a person completes transitional movements between sitting and standing. In this case, the features are:

TABLE 4. Description of the features for each test (BT, WT, and STST), with specified the camera used for their extraction. The joint numbers in the table are shown in Figure 4.

Test	Feature	Description	Side Camera	Frontal Camera
BT	Feet distance	Distance between joints 10 and 13	✓	✓
	Right hand-hip distance	Distance between joints 11 and 7		✓
	Left hand-hip distance	Distance between joints 4 and 8		✓
	Body posture	Distance between the projection of the joint 1 and the midpoint of the segment connecting joints 8 and 11		✓
WT	Feet distance	Distance between joints 10 and 13	✓	✓
	Right hand-hip distance	Distance between joints 11 and 7	✓	
	Left hand-hip distance	Distance between joints 4 and 8	✓	
	Body posture	Distance between the projection of the joint 1 and the midpoint of the segment connecting joints 8 and 11	✓	
STST	Legs/Torso Angle	Angle between the legs and the torso	✓	
	Knee Angle	Angle at knee	✓	
	Right Elbow Angle	Angle at the right elbow		✓
	Left Elbow Angle	Angle at the left elbow		✓

- The angle between the legs and the torso (*(j)* in Figure 4) and the knee angle (*(jj)* in Figure 4). Both angles describe the action of sitting as captured by the side camera.
- The angle at the right (or left) elbow from the frontal camera (*(jjj)* or *(jv)* in Figure 4). These features characterize people's confidence while performing the STST.

To highlight how the defined features represent the different situations that occur when the SPPB tests are performed, Figures 5, 6 and 7 show features plots for each SPPB test and in both cases of one person who performed the test correctly and one who failed. In Figure 5a), for example, the graphs of the hand-hip distances show the poor postural stability of the subject. Significant fluctuations in the

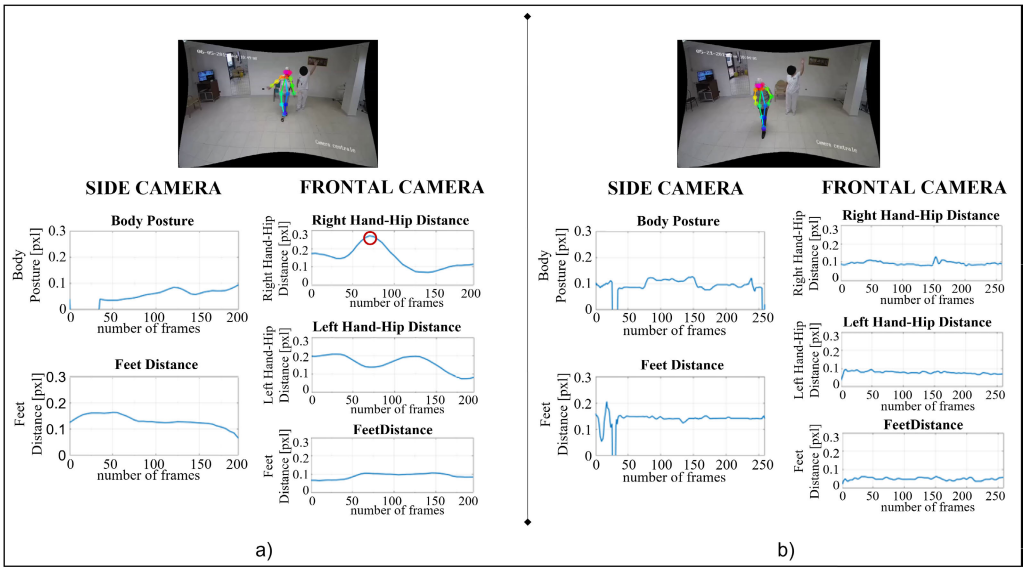

FIGURE 5. Plots of features extracted from the skeletons of two people performing the Tandem position of the BT: a) people of class 0 (unable to maintain balance); b) people of class 3 (correct body posture). The red circle on the feature plot of the Right Hand-Hip Distance indicates the subject's attempt to maintain balance by moving the right arm.

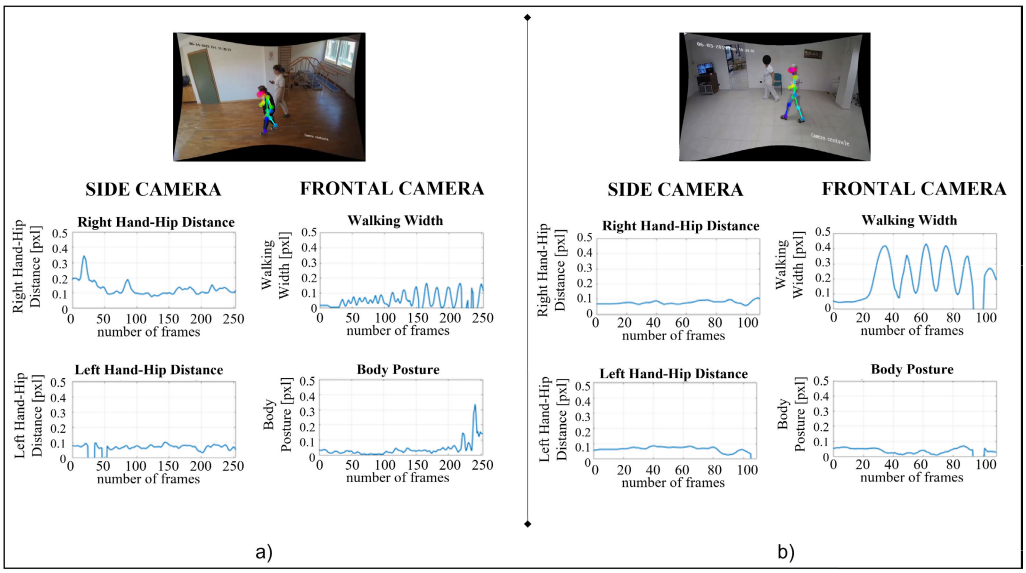

FIGURE 6. Plots of features extracted from the skeletons of two people performing the WT: a) people of class 1 (long execution time); b) people of class 3 (high walking confidence). The graphs of the hand-hip distances and walking widths show the different behavior of the two people in performing the test.

graphs represent the subject's attempts to maintain balance. In contrast, Figure 5b) shows minor fluctuations in the graph, as the subject maintains balance while performing the test.

In the case of WT, in Figure 6a) the feature plots clearly describe a person who needs more time to walk the path, as shown by the sequence of small steps in the walking width graph. On the other hand, the subject in Figure 6b) performs the WT with more confidence, without balancing with the arms.

Finally, in the case of STST, it is evident by the feature plots shown in Figure 7a) how the subject succeeds only two times in standing up. Furthermore, the subject does not keep his arms crossed on the chest, thus failing the test. On the contrary, Figure 7b) shows the case of correct execution of the STST.

B. DEEP NEURAL NETWORK ARCHITECTURES

In this work, the four deep neural networks in Figures 8 and 9 are compared to evaluate the best configuration. The input

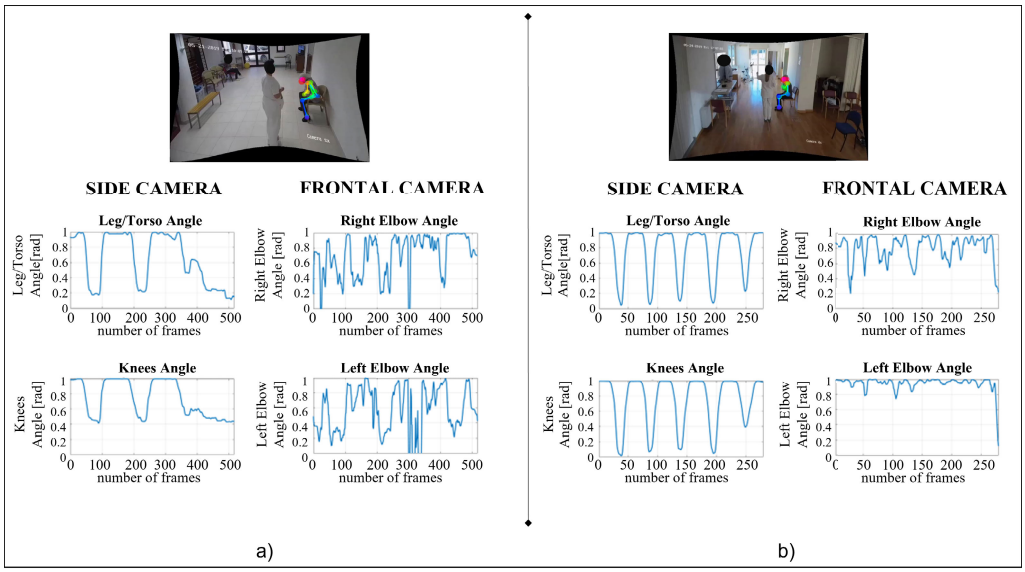

FIGURE 7. Plots of features extracted from the skeletons of two people performing the STST: a) people of class 0 (unable to perform the STST); b) people of class 3 (stands up and sits down 5 times). The left graphs of Leg/Torso and Knee Angles demonstrate that the subject succeeds only two times in standing up. The elbow angles further show the inability to keep the arms crossed on the chest.

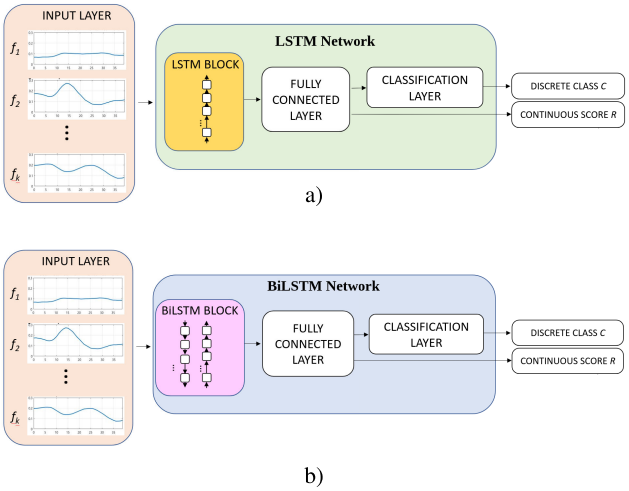

FIGURE 8. Architecture of a) the LSTM network and b) the BiLSTM network.

layer builds the feature vectors by concatenating the features $(f_1, f_2, \ldots, f_k)$ from all the frames contained in the video of the SPPB test, where k depends on the test under examination (see Table 4). Taking into account both side and frontal views, in the case of WT $k = 5$, for STST $k = 4$, while for BT $k = 15$, since BT involves three tests (side-by-side, semi-tandem, and tandem).

The deep network architectures are based on LSTM and BiLSTM models. An LSTM neural network is an extension of a recurring neural network (RNN), suitable for processing time series [45]. Its core is the LSTM block, shown in Figure 8a), which captures essential input features and preserves them over a long period, learning which information is worth storing or erasing through a gating mechanism that controls the memorizing process. In the Bidirectional LSTM (BiLSTM) neural network of Figure 8b), a Backward LSTM and a Forward LSTM cooperate to capture past and future information by letting the data flowing forward and backward [46]. BiLSTM is well-known to achieve better performance than LSTM by modeling the sequences along both directions. In the proposed experiments, both blocks have 100 hidden units

In the proposed work, deep neural networks are designed for two purposes: classification and regression.

- Classification: the input features are processed to select a discrete class C among four classes of interest ($C \in 0, \ldots, 3$). The result is the same as for the physiotherapists, who assign $C = 3$ to successful tests and $C = 0$ to indicate complete inability. The architecture is then completed by a Fully Connected layer, which mixes the information returning a normalized vector, and a Classification layer, which converts the output of the Fully Connected layer into probabilities through a Softmax function and compares them to minimize the cross-entropy.
- Regression: the networks process the input features to produce a continuous score R. This output is strictly dependent on the target class, but, for its nature, can estimate intermediate mobility levels. In this case, the networks are still completed by a Fully Connected layer, whose output is directly interpretable as the final regression result R. During training, the networks try to minimize the half mean-squared error loss, based on the same example of the classification task, i.e. using discrete targets to generalize then and predict continuous scores.

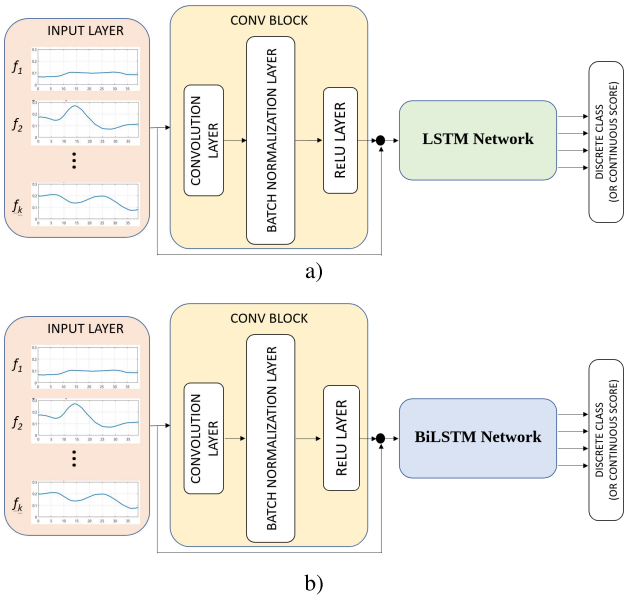

FIGURE 9. Architecture of a) the Conv-LSTM network and b) the Conv-BiLSTM network.

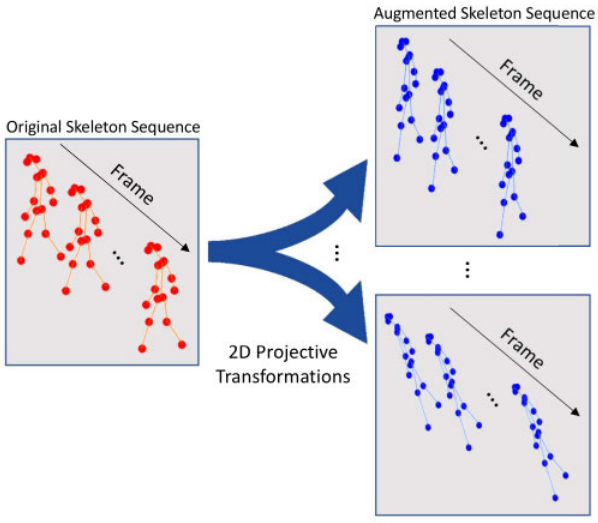

FIGURE 10. Skeleton augmentation process: 2D projective transformations are applied to the original skeletons, obtaining new sequences of augmented skeletons.

To enhance the correlations among features at each time step, we introduce a Convolutional Block (Conv-Block), as shown in Figure 9, obtaining the so-called Conv-LSTM and Conv-BiLSTM networks. The Conv-block consists of a Convolution Layer, a Batch Normalization Layer, and a ReLU Layer, as shown in the yellow box in Figure 9. The Convolution Layer applies several convolutions having $k \times 1$ kernels to the sets of input features at each time step. The Batch Normalization Layer then normalizes the output vectors and is finally rectified using the ReLU function. This block generates a new representation of the input time series to feed the recurrent networks LSTM and BiLSTM of Fig. 8.

C. DATA AUGMENTATION

One of the most frequent problems in machine learning, especially in deep learning, is the lack of a sufficient amount of training data or uneven class balance within the datasets. This problem is even more stringent in this work, where the amount of real data is limited for several reasons (see Section III).

Data augmentation encompasses a suite of techniques that enhance the size and quality of training datasets to build better deep-learning models. In the context of image data, data augmentation includes classical image transformations such as rotation, cropping, zooming, histogram-based methods, color space augmentations, image mixing, and so on [47]. However, these image-based transformations, performed before the skeleton extraction, can induce artifacts in 2D body reconstruction. For this reason, the proposed procedure directly augments the dataset by working on the position of the joints. With reference to Figure 10, data augmentation is made by a set of A rigid geometric transformations of human skeletons, which create different views of the same people, maintaining the relationships between the joints. It is worth noticing that data augmentation is performed after splitting the data into the training and testing sets, to avoid having the augmented features of the same subject in different sets.

Let indicate the joint points in 2D coordinates as $J_p = [x_p, y_p]^T \in \mathbb{R}^2$ in the camera coordinate system, with $p = 0, \cdots, 17$. In general, a point $P = [x, y]^T \in \mathbb{R}^2$ in the 2D Euclidean plane can be described in homogeneous coordinates H as follows [48]:

$$H = \begin{bmatrix} w\,x \\ w\,y \\ w \end{bmatrix} \in \mathbb{P}^2 \qquad w \in \mathbb{R} - \{0\} \tag{1}$$

where $\mathbb{P}^2$ is the 2D projective space defined as $\mathbb{P}^2 = \mathbb{R}^3 - [0,\ 0,\ 0]^T$. For the sake of simplicity, w is typically equal to 1 to have direct transformations between 2D Euclidean and 3D homogeneous coordinates ($P = [x,\ y]^T \leftrightarrow H = [x,\ y,\ 1]^T$).

Let T_i ($i = 1, \cdots, A$), the non-singular 3×3 matrix designed to produce the 2D projective transformation:

$$T_i = \begin{bmatrix} 1 & 0 & E_i \\ 0 & 1 & F_i \\ 0 & 0 & 1 \end{bmatrix} \qquad i = 1, \cdots, A \tag{2}$$

where E_i and F_i are discrete values representing the influence of the vanishing point to the final projection. Large values of E_i and F_i induce close-to-the-origin vanishing points, i.e. parallel lines converging faster. For this reason, these couples of values have been kept small (between 0.001 and 0.01), in the experimental phase, to guarantee reasonable augmentations. Therefore, the new homogeneous 2D coordinates $J'_{p,i}$ of the p-th joint are:

$$J'_{p,i} = \begin{bmatrix} x'_{p,i} \\ y'_{p,i} \\ 1 \end{bmatrix} = T_i \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} \qquad i = 1, \cdots, A \tag{3}$$

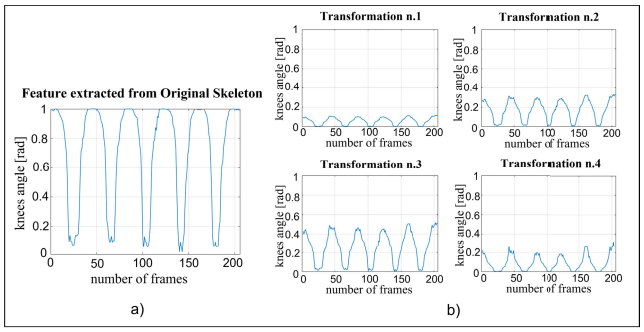

FIGURE 11. a) Plot of a sample feature (knee angle) extracted from the skeleton in an acquired video of STST. b) Different plots of the same feature extracted from the transformed skeletons by applying four different 2D projective transformations.

Since each transformation applies equivalently to all the skeletons, i.e. to all the frames of each video, the size of the resulting dataset after augmentation is A times higher than the initial one in terms of the number of frames. The A parameter has been fixed heuristically by evaluating the performance of the classifiers varying it.

Figure 11 reports the plots of one feature, the knee angle, extracted from the acquired video of a subject performing the STST (Fig. 11a)) and that of four skeletons (Fig 11b)) obtained by applying four different 2D projective transformations. From a first qualitative analysis, the features extracted from the transformed skeletons are still coherent in magnitude and time with the original ones. This aspect is of enormous importance since the features extracted from both original and transformed skeletons differ numerically, ensuring sufficient dataset variability. Still, they refer to the same target class that resembles the same high-level behavior. In the experimental section, we report the classification accuracy, demonstrating the quantitative evidence of the proposed data augmentation procedure.

V. EXPERIMENTS

This section describes the experimental results and details the different data processing steps: data acquisition and classification. All computations have been performed on a 64-bit HP Z840 Workstation, with Intel® Xeon® E5-2699v3 CPU @ 2.30 GHz processor and 256 GB of RAM. To accelerate the training process, all operations have been transferred to a NVIDIA® Quadro® K5200 GPU.

A. DATA ACQUISITION AND PROCESSING

The cameras used for data acquisition are low-cost 4k cameras by HIKVision with 3849×2160 resolution at 20 fps, usually used in video surveillance applications. Due to the dimensions of the gym of the nursing institutes, where videos were acquired, the frontal camera had a focal length of 2.8 *mm*, whereas the side camera had a focal length of 4*mm*. The videos are 246 in total, 74 videos relative to BT, 76 to WT and 96 to STST (see Table 3). Video durations can vary, depending on the test and the participant.

TABLE 5. Accuracy, Precision, and Recall. These quantities are evaluated starting from the computation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy	Precision	Recall
$\frac{TP+TN}{TP+FP+TN+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

Due to the low-level setup of the cameras, the acquired videos presented some limitations, such as a lack of camera synchronization, slightly different camera frame rates, and misalignment of video frames. Therefore, the videos acquired by the side and frontal cameras were first projected on the same timeline, based on the lowest recorded frame rate to improve video uniformity. Then, the couples of videos were manually shifted by as many frames as the delay between the two cameras to achieve synchronization. A signal given by the physiotherapist at the start of each SPPB test was used for this aim. Furthermore, the videos were trimmed to extract only the clips containing the execution of the tests. Finally, a camera calibration procedure was applied to remove image distortion. The OpenPose library is then applied to extract the skeletons. A skeleton tracking procedure has also been developed to detect only the skeleton of the person performing the test, discarding the skeletons of other subjects present in the scene, such as physiotherapists. The first frame of each video is manually labeled by the user to identify who is running the test. Then, for every frame, a Region of Interest (ROI) of 30×30 pixels is selected around each joint of all the subjects in the scene. With an automated process, each joint-related-ROI is compared with the corresponding ones of the subject of interest at the previous frame. Following a voting mechanism, the skeletons of other people in the scenes are discarded, while one of the subjects performing the test is retained.

Then, the obtained dataset of skeletons has been augmented by applying the data augmentation procedure described in Section IV-C.

B. CLASSIFICATION

This section presents the classification results obtained by applying the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM) described in section IV-B. In the following, classifiers will be compared in terms of Accuracy, Precision, and Recall, whose definitions are in Table 5. These metrics are computed by reducing the multi-class problem to multiple binary problems in a *OneVsAll* strategy. Each metric is thus computed four times to assess the classification of each class against the others. The final evaluation metrics are then computed as the arithmetic mean of the four results, weighted by the population of the corresponding class (weighted average).

In the learning phase, the dataset of the extracted features has been divided into training, validation, and test sets. The samples included in the training and validation sets have

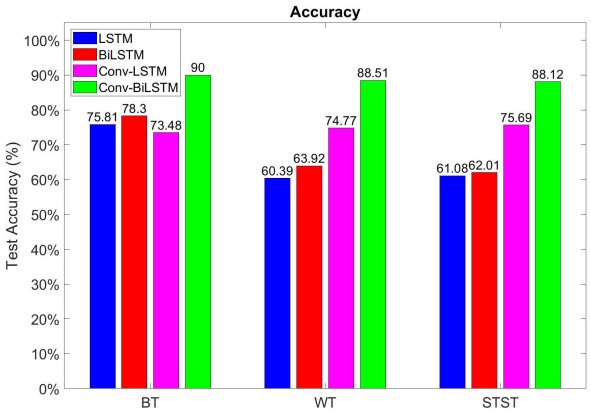

FIGURE 12. Percentages of weighted mean Accuracy of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM and Conv-BiLSTM) for each SPPB test (BT, WT, and STST).

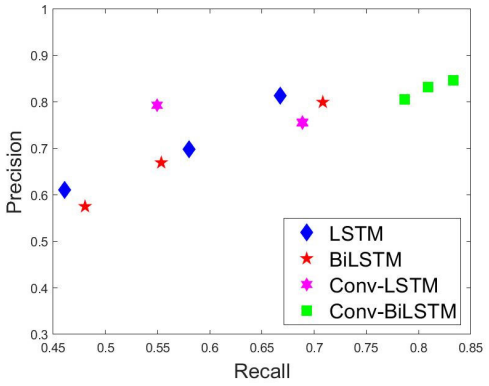

FIGURE 13. Weighted mean values of Precision vs Recall of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM). The three markers for each classifier refer to the three SPPB tests.

been exclusively used for the learning phase. The validation set has been used to assess the model's convergence and stop training when accuracy does not increase for eight consecutive epochs. The test set has been used to evaluate the network's performance in labeling unknown input data. The learning phase results from optimizing a cross-entropy loss function, performed using the Adam optimizer. A 5-fold cross-validation technique has been applied to verify the generalization ability of the networks. For each SPPB test, after cross-validation, only the models with the highest accuracy have been selected to classify elderly people into the four classes (0, 1, 2, 3) defined in Section IV-B.

Figure 12 shows the percentages of weighted mean accuracy of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM), for the three tests BT, WT e STST, respectively. Among the proposed deep architectures, those implementing BiLSTM produce better results than those using LSTM. For example, BiLSTM increases accuracy by an average improvement of 2.31%, considering all three tests of the SPPB. Similarly, the Conv-BiLSTM classifier outperforms the Conv-LSTM one with an average improvement of 14.23%. These results confirm that taking input in forward and backward directions increase the amount of available information, capturing the complex variability of the features. At the same time, introducing the Convolutional Block before the LSTM/BiLSTM networks produces a more significant enhancement of the classification accuracy. In particular, the Conv-LSTM network increases performance in dynamic tests (WT and STST) compared to the results of the LSTM one, as well as the Conv-BiLSTM over the BiLSTM with an average improvement of 19.91%. Indeed, applying convolutional kernels to the input feature vectors transforms the data into new vectors that better characterize the features' spatial correlation, improving the final classification ability. In Figure 13, the weighted averages of Precision vs Recall are reported for each deep neural network architecture and each SPPB test. Precision/Recall metrics also confirm that Conv-BiLSTM architecture outperforms the others.

Additional experiments have been conducted to evaluate how data augmentation affects the classifiers' performance. Figure 14 shows the resulting weighted mean accuracy of the deep classifiers for each SPPB test when A ranges between 0 and 100. At first glance, increasing the dataset size by data augmentation results in an improvement in the average accuracy for any network and any mobility test. However, adding more data leads to longer training time. Consequently, $A = 50$ produces the best trade-off between accuracy and dataset size. It should be noted that Conv-BiLSTM always performs better regardless of the size of the dataset defined by the parameter A.

For the sake of completeness, several machine learning classifiers have been also considered, namely Decision Tree, Naive Bayes, SVM, and KNN classifiers [49]. Also in this case, a 5-fold cross-validation technique has been applied during the learning phase, while the configuration with the maximum accuracy has been selected for the test phase.

Figure 15 shows that the considered traditional machine learning approaches perform worse than the Conv-BiLSTM approach, thus proving the need for a deep model. Only Decision Tree has good accuracy performance for what concerns the BT. In this case, the Decision Tree sets its first levels to find the end of the test, setting close-to-zero thresholds at specific samples of the input feature vectors. Accordingly, the tree classifies the input focusing only on the duration of the test, i.e. how long the subject stands in the same position. The performance of the Decision Tree emphasizes how the duration of the exercise is also an implicit feature that this specific model uses. This quantitative analysis allows for a good accuracy value compared to the other standard models. However, this is still below the best accuracy achievable by deep models, which even consider the quality of execution. This point is much more significant for the WT and the STST, whose classification is much more dependent on the quality of the execution. For this reason, the classification accuracies of WT and STST of the Decision

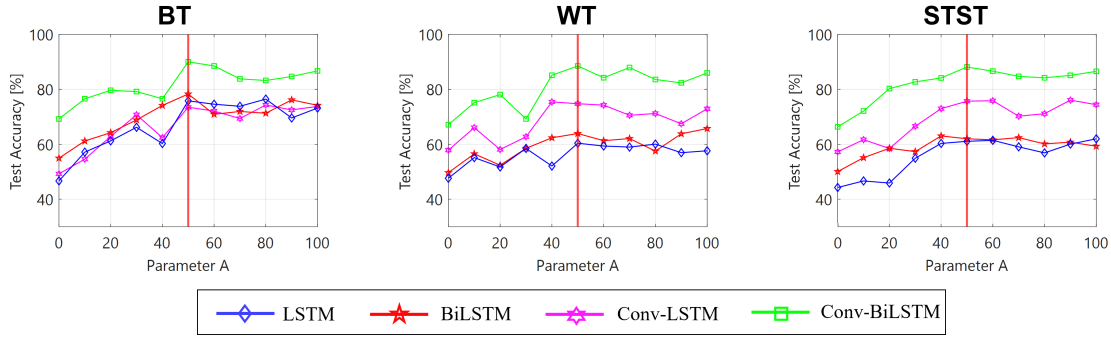

FIGURE 14. Percentages of weighted mean Accuracy of the proposed classifiers varying the *A* parameter, for BT, WT, and STST respectively.

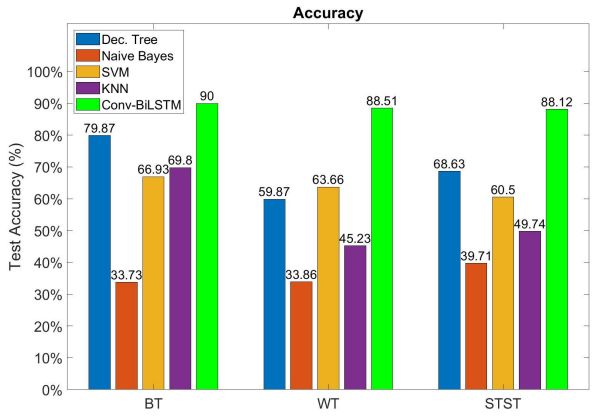

FIGURE 15. Percentages of weighted mean Accuracy of the traditional Machine Learning classifiers compared with the Classification Conv-BiLSTM network for each SPPB test (BT, WT, and STST).

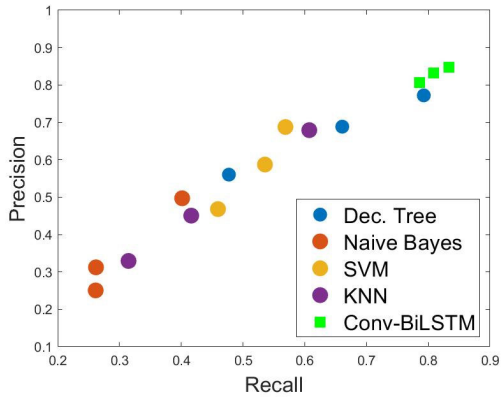

FIGURE 16. Weighted mean values of Precision vs Recall of the traditional neural network architectures (Dec. Tree, Naive Bayes, SVM, KNN) compared with those of the classification model Conv-BiLSTM. The three markers for each classifier refer to the three SPPB tests.

Tree are 28.64% and 19.49% lower than the corresponding values, out of the Conv-BiLSTM model. In Figure 16, the plot of the weighted averages of Precision vs Recall leads to the same conclusion as for the accuracy plot: the Conv-BiLSTM keeps the best performances for the three tests. The Decision Tree classifier has a comparable value only for the BT.

C. REGRESSION

As presented in IV-B, the four deep neural networks have been designed also for regression tasks. To have a proper comparison between Classification and Regression networks, the Root Mean Square Error (RMSE) has been calculated. In classification output, these metrics are computed between discrete integers (expected classes and predicted ones). In contrast, for regression models, they are computed between discrete expected classes and predicted regression values *R*. RMSEs are summarized in Figure 17.

As first remark, the best result of the regression, i.e. the lowest RMSEs, is achieved with the Conv-BiLSTM architecture. This result is in agreement with what has been found for classification, since the use of the preliminary convolutional block can help the BiLSTM network by aggregating features at each frame. At the same time, regression networks always perform worse than their classification counterparts. In principle, this result can be unexpected, as treating the mobility assessment to produce continuous scores rather than discrete ones should prevent heavy misclassifications, e.g. from class 3 to 0, and give results of higher quality. All these considerations would be verified if the training set was actually designed with examples from a regression scenario. However, the initial labeling of the dataset, made by physiotherapists in discrete classes, reduces the ability of regressive networks to create successful models.

D. CONV-BiLSTM CLASSIFIER: IN-DEPTH ANALYSIS

This subsection presents a detailed analysis of the performance of the Conv-BiLSTM network architecture for each class of SPPB tests. These experiments help understand the practical ability of the proposed deep architecture to recognize the classes of people that need particular attention.

Tables 6, 7 and 8 list the Accuracy, Precision, Recall and the resulting weighted averages for each SPPB test and for each class. These results demonstrate that the proposed Conv-BiLSTM, in most cases, can predict the correct class of mobility level for each SPPB test (BT, WT, and STST). With more detail, the weighted mean accuracy is 90% in the case of BT, while it is 88.51% and 88.12% for WT and STST, respectively.

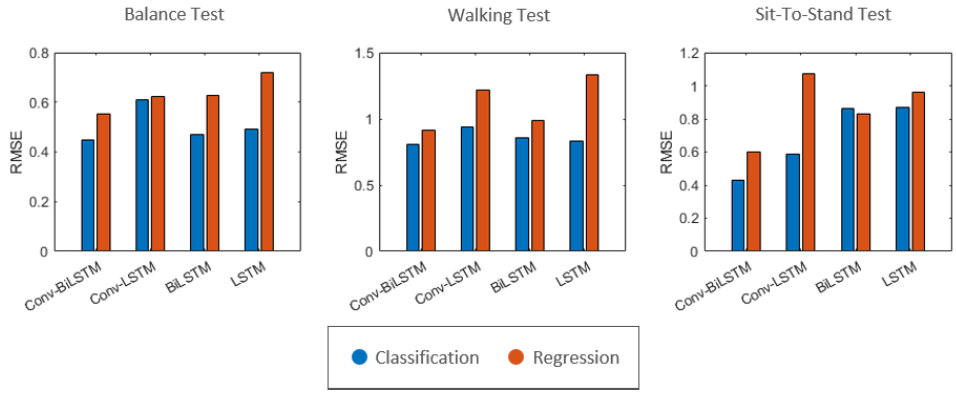

FIGURE 17. Graphs representing the RMSEs values from the Regression Models (orange) vs the Classification Models (blue). Each plot is relative to the exercises within the SPPB test, i.e. BT, WT, and STST.

TABLE 6. Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Balance Test.

BT	Accuracy	Precision	Recall
Class 0	96.92%	87.18%	66.67%
Class 1	90.48%	85.42%	80.39%
Class 2	90.76%	62.68%	87.25%
Class 3	88.52%	90.32%	86.27%
Weighted Mean	90.00%	84.75%	83.33%

TABLE 7. Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Walking Test.

WT	Accuracy	Precision	Recall
Class 0	85.78%	73.51%	97.06%
Class 1	87.38%	93.56%	71.24%
Class 2	94.49%	89.04%	63.73%
Class 3	94.12%	75.47%	78.43%
Weighted Mean	88.51%	83.22%	80.88%

TABLE 8. Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Sit To Stand Test.

STST	Accuracy	Precision	Recall
Class 0	94.53%	73.74%	73.00%
Class 1	88.44%	89.84%	81.95%
Class 2	85.04%	80.38%	69.61%
Class 3	89.27%	60.61%	91.50%
Weighted Mean	88.12%	80.57%	78.64%

Precision, also called positive predictive value, measures how many predictions of a class are true. In our context, it proves the ability of the system to assign the correct mobility level to the person. The weighted averages of Precision are 84.75%, 83.22%, and 80.57%, for the BT, WT, and STST, respectively.

Recall, also known as sensitivity, measures the ability to recognize samples of a specific class. This aspect is fundamental in our experimental context, as it is necessary to be confident of which people need more attention than others. The Recall values in Tables 6, 7 and 8 outline the good performance of the proposed classification model. More precisely, the weighted averages of Recall reach 83.33%, 80.88%, and 78.64% for BT, WT, and STST, respectively.

It is essential to highlight that the obtained results are satisfactory in the particular healthcare context addressed in this work. We have developed a system that makes decisions emulating the decision-making ability of human experts for assessing people's mobility. It is crucial to notice that only specialized physiotherapists with specific competencies can make these evaluations. So developing such an automatic system is of great help for supporting clinicians to identify people with mobility limitations objectively.

Finally, concerning the computational costs, it is straightforward to acknowledge that a longer mobility test leads to longer videos, which require more time for training the corresponding network. In this case, the training time of the architectures for modeling the WT and the STST is higher than that for modeling the BT, although the numbers of training epochs are comparable (26, 28, and 30 for BT, WT, and STST, respectively). The same consideration is still valid for the test phase. The average times for a single video classification are 28 *ms* for the BT, 37 *ms* for the WT, and 49 *ms* for the STST. These last durations are computed on the setup described previously, exploiting the huge capabilities of a GPU implementing Nvidia CUDA drivers. However, the same classifications have been repeated on the single CPU of the same processing unit, leading to average times of 290 *ms* for the BT, 344 *ms* for the WT, and 416 *ms* for the STST. Although CPU processing takes more time than GPU processing, classification times are always much shorter than required for performing every mobility test. This paves the way for future implementations of the trained model on low-resource platforms, such as apps for mobile phones or tablets, towards a fully-integrated telehealthcare system.

VI. DISCUSSION AND CONCLUSION

In recent years, the increase in the elderly population and the need to support diagnostic issues in retirement residences have brought considerable interest in developing telehealthcare systems. This work deals with the complex

problem of the motion ability evaluation of older people. In literature, several automatic systems, both invasive (based on wearable sensors) and noninvasive, have been proposed to measure specific parameters related to gait or posture. On the contrary, few works explore only partially the analysis of people's movement. Currently, the evaluation of motion abilities is carried out by experienced medical personnel who observe people performing some mobility tests through a defined protocol and evaluate their mobility level according to defined rank. Despite the high professionalism of physiotherapists, this evaluation can be affected by their subjectivity, confidence, and experience. Therefore, the development of automatic systems can significantly help medical personnel improve diagnostic accuracy and the elderly themselves by limiting the number of visits to health clinics.

The main contributions of this paper are:

- The feasibility of developing an automated system to assess the motion skills of older people while performing a specific mobility test protocol has been demonstrated. The proposed system is noninvasive for people. It consists of low-cost visual cameras that record videos of people performing the tests and a complete processing framework that extracts significant features and builds models that classify the test executions emulating the complex decision process of physiotherapists.
- The proposed system has been validated using real video data acquired in two nursing institutes hosting elderly people, both healthy and affected by neurodegenerative diseases. Significant features have been extracted from the skeletal representations of the subjects observed. To increase the dataset dimensionality, a data augmentation technique has been applied to the extracted skeletons. Finally, the proposed deep neural network, based on BiLSTM, has been used to classify the observed people's mobility levels. Numerical experiments have been analyzed quantitatively in terms of Accuracy, Precision, and Recall metrics, demonstrating the improvement of results due to preliminary processing made by a convolutional block.
- Several machine learning methods for automatically classifying the motion functionalities of older adults have been compared. Once again, the deep neural network classifier with convolutional filters and a BiLSTM model provides the best performance among all the implemented techniques.
- The proposed deep neural network architectures have also been tuned to perform regression. The results show an improvement in RMSE due to convolutional blocks. However, the input labels (discrete classes) do not constitute a significant dataset for training regression models, which perform worse than the Conv-BiLSTM designed for the classification of patients performing the SPPB test.

From the analysis of the results, we can assert that the proposed approach reaches good performance despite the limited number of video acquisitions. Some research issues will be investigated in deep in future research:

- The acquisition of more consistent datasets regarding the number of observed subjects and the number of assessments will allow a more deep validation of the automatic system. Through the consensus of several expert therapists, class evaluations could overcome the bias of individual assessments.
- One of the main points of this work is the use of real data. The scientific community often complains about the poor availability of data that limits the application and experimentation of machine learning methodologies that need a large amount of data to build significant models and robust systems. Making datasets publicly available can allow the scientific community to compare different approaches, to improve and share statistical analyses.
- The system can be considerably improved with the recently available low-cost and high-performance depth cameras. Currently, the system uses two low-cost RGB cameras. Future developments will involve the use of an appropriately placed RGB-D camera in front of the person performing the mobility tests. The proposed approach remains valid, as the defined features are invariant from the point of view and can be evaluated using depth information. In addition, using a single camera makes the system more flexible, allowing its applicability not only in retirement or nursing homes but also in private homes.

The proposed system reveals the mobility levels of people, supporting clinicians to timely detect mobility anomalies, and preventing dangerous conditions such as falls or worsening health conditions. Furthermore, the development of mobile apps that collect video of people performing mobility tests, extract data, and transmit them to medical staff, could provide good support to increase telehealthcare functionalities. Telehealthcare systems will be a valid instrument for remote monitoring of older adults often unwilling to visit health clinics periodically, reducing time, costs, and efforts.

ACKNOWLEDGMENT

The authors are deeply thankful to Michele Attolico and Giuseppe Bono for their fundamental administrative and technical support.

REFERENCES

[1] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, pp. 1–25, Jul. 2021.

[2] U. Nations, "World population ageing 2020 highlights: Living arrangements of older persons," Dept. Economic Social Affairs, United Nations, New York, NY, USA, Tech. Rep., ST/ESA/SER.A/451, 2020. [Online]. Available: https://www.un.org/development/desa/pd/

[3] *Global Action Plan on the Public Health Response to Dementia 2017–2025*, World Health Organization, Geneva, Switzerland, 2017.

[4] C. Buckley, L. Alcock, R. McArdle, R. Z. U. Rehman, S. D. Din, C. Mazzà, A. J. Yarnall, and L. Rochester, "The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: Insights from gait and postural control," *Brain Sci.*, vol. 9, no. 2, pp. 1–21, 2019.

[5] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. R. D'Orazio, "Human gait analysis in neurodegenerative diseases: A review," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 229–242, Jan. 2022.

[6] R. Soubra, A. Chkeir, and J.-L. Novella, "A systematic review of thirty-one assessment tests to evaluate mobility in older adults," *BioMed Res. Int.*, vol. 2019, pp. 1–17, Jun. 2019.

[7] G. Grossi, R. Lanzarotti, P. Napoletano, N. Noceti, and F. Odone, "Positive technology for elderly well-being: A review," *Pattern Recognit. Lett.*, vol. 137, pp. 61–70, Sep. 2020.

[8] G. Cicirelli, R. Marani, A. Petitti, A. Milella, and T. D'Orazio, "Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population," *Sensors*, vol. 21, no. 10, pp. 1–22, 2021.

[9] L. Brognara, P. Palumbo, B. Grimm, and L. Palmerini, "Assessing gait in parkinson's disease using wearable motion sensors: A systematic review," *Diseases*, vol. 7, no. 1, pp. 1–14, 2019.

[10] 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, Oct. 2017.

[11] A. Cereatti, U. D. Croce, and A. M. Sabatini, *Three-Dimensional Human Kinematic Estimation Using Magneto-Inertial Measurement Units* (Handbook of Human Motion). Cham, Switzerland: Springer, 2017, pp. 1–24.

[12] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports Med.-Open*, vol. 4, no. 1, pp. 1–15, Dec. 2018.

[13] R. Rucco, V. Agosti, F. Jacini, P. Sorrentino, P. Varriale, M. D. Stefano, G. Milan, P. Montella, and G. Sorrentino, "Spatio-temporal and kinematic gait analysis in patients with frontotemporal dementia and Alzheimer's disease through 3D motion capture," *Gait Posture*, vol. 52, pp. 312–317, Feb. 2017.

[14] N. K. Mangal and A. K. Tiwari, "A review of the evolution of scientific literature on technology-assisted approaches using RGB-D sensors for musculoskeletal health monitoring," *Comput. Biol. Med.*, vol. 132, pp. 1–15, May 2021.

[15] F. Wang, E. Stone, M. Skubic, J. M. Keller, C. Abbott, and M. Rantz, "Toward a passive low-cost in-home gait assessment system for older adults," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 2, pp. 346–355, Mar. 2013.

[16] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, "A review of 3D human pose estimation algorithms for markerless motion capture," *Comput. Vis. Image Understand.*, vol. 212, Nov. 2021, Art. no. 103275.

[17] C. P. Hensley, D. Millican, N. Hamilton, A. Yang, J. Lee, and A. H. Chang, "Video-based motion analysis use: A national survey of orthopedic physical therapists," *Phys. Therapy*, vol. 100, no. 10, pp. 1759–1770, Sep. 2020.

[18] L. Romeo, R. Marani, N. Lorusso, M. T. Angelillo, and G. Cicirelli, "Vision-based assessment of balance control in elderly people," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–6.

[19] N. Kour and S. Arora, "Computer-vision based diagnosis of Parkinson's disease via gait: A survey," *IEEE Access*, vol. 7, pp. 156620–156645, 2019.

[20] R. A. Clark, B. F. Mentiplay, E. Hough, and Y. H. Pua, "Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives," *Gait Posture*, vol. 68, pp. 193–200, Feb. 2019.

[21] P. F. Dajime, H. Smith, and Y. Zhang, "Automated classification of movement quality using the Microsoft Kinect V2 sensor," *Comput. Biol. Med.*, vol. 125, Oct. 2020, Art. no. 104021.

[22] A. Ejupi, M. Brodie, Y. J. Gschwind, S. R. Lord, W. L. Zagler, and K. Delbaere, "Kinect-based five-times-sit-to-stand test for clinical and in-home assessment of fall risk in older people," *Gerontology*, vol. 62, no. 1, pp. 118–124, 2016.

[23] O. Mazumder, S. Tripathy, S. Roy, S. Chakravarty, D. Chatterjee, and A. Sinha, "Postural sway based geriatric fall risk assessment using Kinect," in *Proc. IEEE Sensors*, Glasgow, U.K., Nov. 2017, pp. 1–3.

[24] F. Romano, P. Colagiorgio, A. Buizza, F. Sardi, and S. Ramat, "Extraction of traditional COP-based features from COM sway in postural stability evaluation," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 3715–3718.

[25] O. Tupa, A. Procházka, O. Vysata, M. Schätz, J. Mares, M. Valis, and V. Marík, "Motion tracking and gait feature estimation for recognising Parkinson's disease using MS Kinect," *Biomed. Eng. OnLine*, vol. 14, no. 1, pp. 1–20, Dec. 2015.

[26] J. Medina-Quero, C. Shewell, I. Cleland, J. Rafferty, C. Nugent, and M. E. Estevez, "Computer vision-based gait velocity from non-obtrusive thermal vision sensors," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 391–396.

[27] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, R. J. Valdivieso-Sarabia, J. Mora-Pascual, and J. M. García-Chamizo, "Gait analysis using computer vision based on cloud platform and mobile device," *Mobile Inf. Syst.*, vol. 2018, pp. 1–10, Jan. 2018.

[28] L. Dranca, L. D. A. R. De Mendarozketa, A. Goñi, A. Illarramendi, I. N. Gomez, M. Delgado Alvarado, and M. C. Rodríguez-Oroz, "Using Kinect to classify Parkinson's disease stages related to severity of gait impairment," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–15, Dec. 2018.

[29] V. Dentamaro, D. Impedovo, and G. Pirlo, "Gait analysis for early neurodegenerative diseases classification through the kinematic theory of rapid human movements," *IEEE Access*, vol. 8, pp. 193966–193980, 2020.

[30] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Dec. 2017.

[31] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.

[32] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[33] S. Zhang, S. K. Poon, K. Vuong, A. Sneddon, and C. T. Loy, *A Deep Learning-Based Approach for Gait Analysis in Huntington Disease* (Studies in Health Technology and Informatics), vol. 264. Amsterdam, The Netherlands: IOS Press, 2019, pp. 477–481.

[34] S. Bringas, S. Salomón, R. Duque, J. L. Montana, and C. Lage, "A convolutional neural network-based method for human movement patterns classification in Alzheimer's disease," *Multidisciplinary Digit. Publishing Inst. Proc.*, vol. 31, no. 1, p. 72, 2019.

[35] F. Luna-Perejon, M. J. Dominguez-Morales, and A. Civit-Balcells, "Wearable fall detector using recurrent neural networks," *Sensors*, vol. 19, no. 22, pp. 1–18, 2019.

[36] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks*. Berlin, Germany: Springer, 2012.

[37] C. Tunca, G. Salur, and C. Ersoy, "Deep learning for fall risk assessment with inertial sensors: Utilizing domain knowledge in spatio-temporal gait parameters," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1994–2005, Jul. 2020.

[38] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1110–1118, Mar. 2021.

[39] K. Oh and K. Jung, "GPU implementation of neural networks," *Pattern Recognit.*, vol. 37, pp. 1311–1314, Jun. 2004.

[40] Z. Meng, M. Zhang, C. Guo, Q. Fan, H. Zhang, N. Gao, and Z. Zhang, "Recent progress in sensing and computing techniques for human activity recognition and motion analysis," *Electronics*, vol. 9, no. 9, pp. 1–19, 2020.

[41] J. Fish, *Short Physical Performance Battery* (Encyclopedia of Clinical Neuropsychology). Berlin, Germany: Springer, 2011, pp. 2289–2291.

[42] L. Romeo, R. Marani, A. Petitti, A. Milella, T. D'Orazio, and G. Cicirelli, "Image-based mobility assessment in elderly people from low-cost systems of cameras: A skeletal dataset for experimental evaluations," in *Ad-Hoc, Mobile, and Wireless Networks*. Berlin, Germany: Springer, 2020, pp. 125–130.

[43] L. Romeo, R. Marani, T. D'Orazio, and G. Cicirelli. (2020). *SPPBdataset*. [Online]. Available: https://github.com/ispstiima/SPPBdataset

[44] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[46] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 436–440.
[47] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[48] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[49] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

LAURA ROMEO received the B.E. and M.S. degrees (cum laude) in automation engineering from the Polytechnic University of Bari, Italy, in 2017 and 2019, respectively. She is currently pursuing the Ph.D. degree. Since October 2019, she has been working a Research Fellow with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA), National Research Council of Italy (CNR). In November 2020, she has been accepted as a Ph.D. Student in industry 4.0 at the Polytechnic of Bari, Italy, for the project "IMPACT: vIsion devices and systems for Monitoring the wellbeing of oPerators in a workspAce shared with Cobots in indusTry 4.0." She is also an Automation Engineer.

ROBERTO MARANI received the B.E., M.S., and Ph.D. degrees, in 2006, 2008, and 2012, respectively. Since 2012, he has been working with CNR-STIIMA, where he cooperates with the Intelligent Sensing and Perception Group. Since July 2018, he has been a Permanent Staff Researcher. He is currently an Electronics Engineer. His studies in electronics, optics, computer vision, machine learning, pattern recognition, and deep learning affect several application fields, ranging from quality control in industrial production processes, precision agriculture, robotics, geology, archaeology, reverse engineering, and video surveillance. He is the author of more than 100 scientific articles in international peer-reviewed journals and 40 proceedings of international conferences. His research interest includes design and implementation of hardware and software for autonomous intelligent systems.

TIZIANA D'ORAZIO received the degree (summa cum laude) in computer science from the University of Bari, Italy. Since 1997, she has been working with the Institute of Signal and Image Processing, Italian National Research Council (CNR). She is currently a Senior Researcher with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA), CNR. She is the author of more than 150 articles published in international journals and international conference proceedings and coauthored three international patents on visual systems for event detection in sportive contexts. Her research interests include pattern recognition, artificial intelligence, image and signal processing for robotic application, and intelligent perception systems.

GRAZIA CICIRELLI received the Laurea degree (magna cum laude) in computer science from the University of Bari, in 1994. From 1995 to 2001, she was a Fellow Researcher of the National Research Council of Italy (CNR), working on activities related to robotics and image processing. Since 2001, she has been a Permanent Staff Technologist Researcher with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA), CNR. She is the author of numerous research papers published in international conference proceedings, national and international journals. Her research interests include pattern recognition, artificial intelligence, machine learning, image processing for robotic applications, and intelligent systems for visual surveillance.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE' within the CRUI CARE Agreement