Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson’s disease: A systematic review

Expert Systems With Applications 229 (2023) 120541

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review

Luis Sigcha [a,b,*], Luigi Borzì [c], Federica Amato [c], Irene Rechichi [c], Carlos Ramos-Romero [d], Andrés Cárdenas [e], Luis Gascó [f], Gabriella Olmo [c]

[a] ALGORITMI Research Center, School of Engineering, University of Minho, 4800-058 Guimarães, Portugal
[b] Data-Driven Computer Engineering (D^2iCE) Group, Department of Electronic and Computer Engineering, University of Limerick, Limerick, V94 T9PX, Ireland
[c] Data Analytics and Technologies for Health (ANTHEA) Lab, Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy
[d] Acoustics Research Centre, University of Salford, M5 4WT Manchester, UK
[e] I2CAT Foundation, Barcelona, Spain
[f] Barcelona Supercomputing Center, 08034 Barcelona, Spain

ARTICLE INFO

Keywords:
Parkinson's disease
Motor symptoms
Non-motor symptoms
Wearables
Body-worn sensors
Machine learning
Deep learning
Neural networks
Convolutional neural networks
Recurrent neural networks

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that produces both motor and non-motor complications, degrading the quality of life of PD patients. Over the past two decades, the use of wearable devices in combination with machine learning algorithms has provided promising methods for more objective and continuous monitoring of PD. Recent advances in artificial intelligence have provided new methods and algorithms for data analysis, such as deep learning (DL). The aim of this article is to provide a comprehensive review of current applications where DL algorithms are employed for the assessment of motor and non-motor manifestations (NMM) using data collected via wearable sensors. This paper provides the reader with a summary of the current applications of DL and wearable devices for the diagnosis, prognosis, and monitoring of PD, in the hope of improving the adoption, applicability, and impact of both technologies as support tools. Following PRISMA (Systematic Reviews and Meta-Analyses) guidelines, sixty-nine studies were selected and analyzed. For each study, information on sample size, sensor configuration, DL approaches, validation methods and results according to the specific symptom under study were extracted and summarized. Furthermore, quality assessment was conducted according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) method. The majority of studies (74%) were published within the last three years, demonstrating the increasing focus on wearable technology and DL approaches for PD assessment. However, most papers focused on monitoring (59%) and computer-assisted diagnosis (37%), while few papers attempted to predict treatment response. Motor symptoms (86%) were treated much more frequently than NMM (14%). Inertial sensors were the most commonly used technology, followed by force sensors and microphones. Finally, convolutional neural networks (52%) were preferred to other DL approaches, while extracted features (38%) and raw data (37%) were similarly used as input for DL models. The results of this review highlight several challenges related to the use of wearable technology and DL methods in the assessment of PD, despite the advantages this technology could bring in the development and implementation of automated systems for PD assessment.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder caused by the loss of dopaminergic neurons in the region of the substantia nigra pars compacta, located in the midbrain. PD manifests with various movement-related symptoms (e.g., tremor, rigidity, bradykinesia, akinesia, or postural instability) termed motor symptoms and mental health-related symptoms (e.g., memory problems or dementia) termed non-motor manifestations (NMM) (Armstrong & Okun, 2020; Goetz, 2011; Tolosa, Wenning, & Poewe, 2006). In general, symptoms appear gradually and become more evident as the disease progresses (Sica et al., 2021). The presence of these symptoms leads to a gradual loss of

* Corresponding author at: ALGORITMI Research Center, School of Engineering, University of Minho, 4800-058 Guimarães, Portugal.
E-mail addresses: luis.sigcha@ul.ie (L. Sigcha), luigi.borzi@polito.it (L. Borzì), federica.amato@polito.it (F. Amato), irene.rechichi@polito.it (I. Rechichi), c.a.ramosromero@salford.ac.uk (C. Ramos-Romero), andres.cardenas@i2cat.net (A. Cárdenas), lgasco@bsc.es (L. Gascó), gabriella.olmo@polito.it (G. Olmo).

https://doi.org/10.1016/j.eswa.2023.120541
Received 2 January 2023; Received in revised form 22 May 2023; Accepted 23 May 2023
Available online 27 May 2023

autonomy, resulting in a reduced quality of life (QoL) for patients (Zhao et al., 2021).

The prevalence of PD increases with age, being rare before the age of 50 and more common in men than in women (Reeve, Simcox, & Turnbull, 2014; Wirdefeldt, Adami, Cole, Trichopoulos, & Mandel, 2011). Due to the ageing population, the prevalence of the disease is expected to increase significantly worldwide, from 6.9 million in 2015 to approximately 12 million in 2040 (Dorsey, Sherer, Okun, & Bloem, 2018).

The diagnosis of PD is based on clinical criteria and the presence of bradykinesia, described as slowness of movements, together with symptoms such as rigidity and/or tremor and the presence of supportive features (Jankovic, 2008; Kobylecki, 2020). Despite advances in the fields of neuroimaging, genetics, and biomarkers, the accuracy of clinical diagnosis of PD has been suboptimal, with no substantial improvement, especially in the early stages of the disease or in the presence of signs of atypical parkinsonism (Rizzo et al., 2016).

Currently, PD has no cure, and pharmacotherapy and surgery are the most commonly used supporting therapies. Drugs such as levodopa and dopamine agonists remain the most effective treatments to improve the signs of PD (Armstrong & Okun, 2020; Reich & Savitt, 2019). These drugs offer good control of motor symptoms in the early stages of the disease but do not stop neurodegeneration, disease progression, or increased disability (Balestrino & Schapira, 2020). Moreover, after several years of drug treatment, the efficacy of these therapies decreases, and side effects such as motor fluctuations and dyskinesias (e.g., involuntary movements) occur (Jankovic, 2005).

The current standard for the assessment of PD is the clinical examination of patients by a neurology specialist using semi-quantitative rating scales. The most commonly used scale to measure PD progression is the Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (Goetz et al., 2008). During the examination, the neurologist assesses the patient's motor and mental state while performing specific tasks.

Although clinical scales such as the MDS-UPDRS are widely used, they are affected by subjectivity and often high inter-rater and intra-rater variability (Monje, Foffani, Obeso, & Sánchez-Ferro, 2019). Moreover, clinical assessments are commonly performed during pre-scheduled medical visits (e.g., every 6 or 9 months) (Albanese, 2013). In this context, the limited periodicity of the visit makes it difficult for neurologists to assess short-term changes in the patient's disability, while the subjective nature of the clinical examination may lead to biased assessment results (Rodríguez-Martín et al., 2022). This makes it difficult to implement appropriate therapeutic adjustments, thus reducing the effectiveness of treatments (Bhidayasiri & Martinez-Martin, 2017). Therefore, the need for objective assessment mechanisms has led to the emergence of technological tools to facilitate PD management and optimize long-term follow-up (Channa, Popescu, & Ciobanu, 2020; Luis-Martínez, Monje, Antonini, Sánchez-Ferro, & Mestre, 2020; Monje et al., 2019; Warmerdam et al., 2020).

Currently, the combination of technologies such as wearable devices (wearables) and artificial intelligence offers several possibilities to monitor patients with chronic diseases (Luis-Martínez et al., 2020). In particular, the use of these technologies in PD can help overcome the drawbacks of monitoring, such as the lack of objective information obtained by traditional methods or the fact that the hospital environment is not considered the most convenient scenario for a proper patient assessment (Rodríguez-Martín et al., 2022). In this context, continuous monitoring in daily living conditions becomes the most appropriate way to characterize the patient's condition.

1.1. Review objectives

This article aims to provide a comprehensive review of current applications in which deep learning (DL) architectures and wearables are used to evaluate motor aspects and NMM. Specifically, this review summarizes the application areas, data processing approaches, data sets used to evaluate the performance of the algorithms, and experimental methodologies proposed for automatic PD assessment. This article provides quick access to relevant literature and allows for the identification of existing gaps and opportunities for the application of DL techniques and algorithms to improve PD outcomes and management.

Previous studies have systematically and extensively examined the use of machine learning (ML) and/or wearables for the assessment of PD. Channa et al. (2020) and Mughal, Javed, Rizwan, Almadhor, and Kryvinska (2022) examined the state-of-the-art (SOTA) of wearable solutions to improve early diagnosis and monitoring of PD. However, ML algorithms were not the main focus. Loh et al. (2021), Mei, Desrosiers, and Frasnelli (2021), and Tanveer, Rashid, Kumar, and Balasubramanian (2022) analyzed studies that proposed ML and DL models for the automatic assessment of PD. Different types of modalities were considered, including imaging techniques and wearable sensors. However, these studies focused specifically on computer-assisted diagnosis and/or differential diagnosis tools. Consequently, only a few PD symptoms, such as gait, writing, and speech, were evaluated. In addition, little information on wearables (e.g., type, number, location) was provided. Finally, Giannakopoulou, Roussaki, and Demestichas (2022) reviewed research works that used ML in combination with wearable and non-wearable sensors to monitor PD. However, information was reported according to the type of sensor, which makes it difficult to compare different approaches for assessing PD symptoms.

This article aims to complement the current literature and provide the reader with a summary of the current applications of DL and wearables for the diagnosis, prognosis, and monitoring of motor manifestations and NMMs, in the hope of improving the adoption, applicability, and impact of both technologies as support tools in PD. The present study reviews relevant research articles published between 2012 and 2022 focused on the diagnosis, monitoring, and prediction of the response to treatment using DL approaches and wearables. The main contributions of this paper are summarized as follows.

a. The use of wearable devices is summarized, including information on the type and number of sensors and their placement on the human body. DL methods, data processing and transformation techniques, and objectives (classification, regression) are discussed. Data sources, sample sizes, and data collection procedures are presented, as well as performance evaluation strategies and metrics.

b. The information extracted is reported according to the specific motor symptoms and NMMs examined. Data, wearable sensors, data processing, DL methods, validation procedures, and results are reported for each symptom. This allows an immediate comparison between the different proposed solutions.

c. Limitations and risk of bias in the reviewed literature are presented. Based on these considerations, some guidelines are given, in the hope of improving the future application of DL and wearables for comprehensive PD assessment and monitoring.

The rest of the document is organized as follows. Section 2 presents the context, including a description of the wearable technology and the ML/DL algorithms. Section 3 describes the methodology used for the systematic review, including the search strategy used to find relevant articles. Section 4 presents the results of this literature review, summarizing wearable sensors and DL approaches. In addition, this section includes a comprehensive review of the specific symptoms to which these technologies have been applied. Section 5 provides a discussion of the main findings, potential, and limitations of DL and wearable technology as a tool to support PD assessment. Finally, the conclusions of this work are given in Section 6.

2. Background

This section provides an overview of the wearable devices used for healthcare applications, further discussing the wearable sensors used for the diagnosis and monitoring of PD (Section 2.1). In addition, classic ML approaches used for data analysis are described, together with more advanced DL methods (Section 2.1). Finally, the performance evaluation strategies and metrics are summarized in Section 2.3.

2.1. Wearable devices

Wearables is the common term to describe electronic devices that can be worn on the body, integrated into clothing, or designed as wearable accessories (Bonato, 2010; Johansson, Malmgren, & Alt Murphy, 2018). This group of devices includes smart glasses, smartphones, smart watches, smart clothes, and smart shoes, among others.

In recent years, wearable devices have shown the potential to address some of the limitations of conventional assessment in healthcare and medicine through digital and mobile health (m-health) technologies. They enable continuous, longitudinal health monitoring outside the clinic, in a discreet and comfortable manner (Dunn, Runge, & Snyder, 2018).

Wearable devices are equipped with special hardware and software technology to collect accurate measurements of physical parameters. In addition, these devices offer excellent wireless communication, processing, and data storage capabilities (Heikenfeld et al., 2018; Park, Lee, & Park, 2019). The sensors embedded in wearables are as follows:

- *Inertial sensors* (i.e. gyroscopes, accelerometers, and magnetometers) can provide spatio-temporal kinematic parameters.
- *Acoustic sensors* such as microphones can be used to detect speech disturbances from voice signals.
- *Optical sensors* use light to detect biological signals such as heart rate and oxygen saturation.
- *Electrical sensors* such as electrocardiography (ECG), electroencephalography (EEG), and electromyography (EMG) measure heart rate, brain activity, and muscle movement, respectively.

In the context of monitoring neurological disorders, wearables offer the possibility to perform objective and long-term monitoring of movement patterns or physiological variables in laboratory, hospital, and free-living environments (Johansson et al., 2018; Mughal et al., 2022; Rovini, Maremmani, & Cavallo, 2017). In this sense, the proper implementation of this technology has the potential to provide accurate assessment of motor symptoms (Channa et al., 2020; Del Din, Kirk, Yarnall, Rochester, & Hausdorff, 2021; Rovini et al., 2017) and NMM (Morgan et al., 2020; van Wamelen et al., 2021), resulting in improved diagnosis, more sensitive monitoring, and more precise adjustments of drug therapies (Monje et al., 2019).

Despite the potential of wearable devices to optimize the clinical management of PD, this technology has not currently achieved widespread clinical adoption due to several factors, including a lack of technical and clinical validation. This situation hinders obtaining approval from regulatory bodies (e.g., European Medicines Agency, Food and Drug Administration) (Espay et al., 2019). Furthermore, the lack of validation of the algorithms used to assess PD under real-world conditions was identified, partly due to the lack of gold-standard references to compare against (Del Din et al., 2021).

2.2. Machine and deep learning

Despite the potential of wearables for cost-effective data collection, the data generated by these devices must be processed to derive clinically relevant information. The management of a huge amount of data (Big Data) has been achieved with the use of artificial intelligence and data analysis techniques such as data mining.

For this reason, ML has become a key element in the development of remote monitoring systems based on wearables. ML algorithms offer methods for analyzing sensor data and extracting meaningful information or discovering hidden patterns in a semi-automatic way. For PD assessment, ML techniques have been successfully applied to monitor various motor symptoms and NMM using motion data, video, neuroimaging, voice, and cerebrospinal fluid, among others (Lu et al., 2020; Mei et al., 2021; Morgan et al., 2020; van Wamelen et al., 2021).

In traditional (shallow) ML approaches, a process of feature extraction and selection is often necessary, as ML models alone are unable to learn from high-dimensional data in their raw form (e.g., medical images or time series recorded by sensors) (Mirza et al., 2019). The complexity in the design and selection of feature sets is an obstacle to the implementation of large-scale monitoring systems and limits the implementation of ML in clinical applications. However, among ML techniques, DL algorithms have been successfully employed in recent years in applications where shallow ML algorithms were traditionally used, leading to state-of-the-art (SOTA) applications for PD management (Mei et al., 2021; Tanveer et al., 2022)

One of the main advantages of using DL over shallow ML is that the former can extract high-level features directly from the data. Therefore, DL approaches allow the development of end-to-end models that reduce the time and effort required for the design of classical pipeline-based approaches, including the selection of appropriate features (Alzubaidi et al., 2021; Sarker, 2021).

DL is a representation learning method based on complex networks. These are composed of several processing layers that can learn data representations with multiple levels of abstraction (LeCun, Bengio, & Hinton, 2015). A DL architecture can be defined as an artificial neural network (ANN) with two or more hidden layers (Shamshirband, Fathi, Dehzangi, Chronopoulos, & Alinejad-Rokny, 2021). Although conventional ANN consists of at least three main components, i.e. the input layer, the hidden layer, and the output layer, architectures known as deep neural networks (DNNs) use many more hidden layers to enhance the predictive capabilities of the network (Deng & Yu, 2014). In a conventional DNN, input values are weighted, bias-corrected, and passed through a non-linear activation function, such as rectified linear unit (ReLU) or softmax, to obtain an output (Schmidhuber, 2015).

DL offers a wide variety of architectures, including convolutional neural networks (CNN), recurrent neural networks (RNN), and Transformer networks. These architectures employ different processing layers, such as densely connected (dense), convolutional, attention, or recurrent, including gated recurrent unit (GRU) or long short-term memory (LSTM). When several of these processing layers are stacked as intermediate layers, highly abstract functions can be created, able to automatically extract high-level features from the raw data.

2.3. Performance evaluation

When developing ML models for classification or regression purposes, the evaluation of model performance is of paramount importance. To perform this procedure correctly, the original data set is divided into training and test sets. The former is used to train the model and optimize its architecture and internal parameters. The optimized model is then tested on the test subset, and the performance is evaluated on the latter. There are different types of validation/testing approaches, based on the criterion of splitting data between the training and the test set.

The leave-one-subject-out (LOSO) validation consists of training the classification model with data from all but one patient, which is used as a test. This process is repeated N times, with N corresponding to the total number of subjects. This is a subject-independent (S-I) validation procedure, as data from a single patient are not shared between training and test subsets, thus providing generalized models (i.e., models with a robust ability to generalize to unseen subjects).

Hold-out validation/testing consists of using a certain percentage of data for training and the rest for testing. K-fold cross-validation (CV) consists of dividing the data set into k parts, then iteratively training the classification model using k-1 parts and testing the remaining part. The procedure is performed k times. Both hold-out and k-fold validation can be performed with a subject-independent or subject-dependent (S-D) approach. In the latter case, models are trained and tested on data from a single subject to develop subject-specific models that can work well on a single patient. Finally, the random-shuffle (RS) strategy consists of randomly shuffling the original data set before validation/testing by mixing data from different patients together. This

procedure does not guarantee the independence of the subjects in the training and test subsets. Therefore, it may produce overestimated results and lead to overfitting, thus reducing the generalization capability of the algorithm.

To evaluate classification performance, the following metrics are defined. True positives (TP) are true samples correctly identified by the model. False positives (FP) represent negative samples incorrectly predicted as positive. False negatives (FN) correspond to positive samples that were not detected by the model. Finally, true negatives (TN) represent correctly classified negative instances. Sensitivity/recall and specificity (Eq. (1)) represent the algorithms' capability to detect TP and TN samples, respectively. Accuracy (Eq. (2)) summarizes the performance of a classification model as the percentage of correct predictions. The F-score is the harmonic mean of sensitivity and precision (Eq. (3)), with precision computed as in Eq. (4). In the case of unbalanced data, the F-score is preferred to accuracy as the global classification metric. Finally, the area under the receiver operating characteristic (AUROC) measures the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve, while the equal error rate (EER) corresponds to the error observed at the point on the ROC curve where sensitivity equals specificity.

$$Sensitivity = \frac{TP}{TP+FN} \qquad Specificity = \frac{TN}{TN+FP} \tag{1}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$F-score = \frac{2 \cdot Sensitivity \cdot Precision}{Sensitivity + Precision} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Regression metrics are calculated comparing predictions (y) with the true values ($\hat{y}$). Specifically, Pearson's correlation coefficient (Eq. (5)) measures how well the model fits the dependent variable, i.e. how much variability in the dependent variable can be explained by the model; it ranges between 0 and 1, with larger values indicating better performance. Root mean square error (RMSE, Eq. (6)) and mean absolute error (MAE, Eq. (7)) are absolute measures of the goodness of fit, providing the entity of deviation from the target values. While MAE treats all errors the same, RMSE gives a greater penalization to large prediction errors.

$$r = \sqrt{1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})}} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}| \tag{7}$$

3. Methods

This section presents the methods used to conduct the systematic review. This review was registered in the Prospective International Register of Systematic Reviews (PROSPERO ID: CRD42021283099), a database of systematic review protocols maintained by the Centre for Reviews and Dissemination at the University of York. Furthermore, the methodology used in this review follows the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher, Liberati, Tetzlaff, & Altman, 2009) as a reporting guideline. The remainder of this section describes the Research questions (Section 3.1), the search strategy (Section 3.2), inclusion and exclusion criteria (Section 3.3), data extraction (Section 3.4), and the criteria used to perform the quality report assessment of the selected studies (Section 3.5).

3.1. Research questions

This review aims to collect, process, and analyze studies in the literature to gather information on DL methods applied to wearable sensor data for motor symptoms and NMM assessment. To achieve this goal, the following research questions were raised.

RQ-1 What are the motor and NMM in which DL methods are used?
RQ-2 What wearable devices are used for the evaluation of PD?
RQ-3 What DL algorithms have been used to assess motor and NMM in PD?
RQ-4 What are the challenges and opportunities present in the current use of DL to assess motor and NMM in PD?

3.2. Search strategy

On April 15, 2022, a literature search was conducted on PubMed, IEEE Xplore, and Web of Science databases for all the returned results. The string included keywords related to the disease under investigation, the sensors used to collect the data, and the DL algorithms used to process the data. Specifically, the Boolean search string used was as follows:

(Parkinson) **AND** *(deep learning OR neural network OR deep neural network OR convolutional neural network OR CNN OR recurrent neural network OR RNN OR long short-term memory OR LSTM OR autoencoder OR AE OR deep belief network OR DBN)* **AND** *(sensor OR wearable OR body-worn OR clothes OR internet of things OR inertial OR IMU OR accelerometer OR gyroscope OR smartphone OR smartwatch OR actigraphy OR force sensor OR insoles OR electromyography OR EMG OR electroencephalography OR EEG OR polysomnography OR PSG OR electrooculogram OR EOG OR electrocardiography OR ECG OR skin conductance OR GSR)*.

No additional filters were applied in the literature search. All retrieved studies were systematically identified and screened, and the data were extracted for relevant information following the PRISMA guidelines (Moher et al., 2009).

3.3. Inclusion and exclusion criteria

The topic of this review concerns the diagnosis and monitoring of PD, along with the prediction of the treatment response. In this study, journal articles published between January 2012 and April 2022 and written in English were included if they used DL methods on data extracted from wearables for diagnosis or monitoring of PD. The exclusion criteria were as follows:

1. The literature was not written in English.
2. Papers without peer review, conference papers, books, book chapters, or published as "letter", "comments", "case reports", "surveys" or "reviews".
3. Studies related to Parkinsonism and/or diseases other than PD.
4. Studies that did not use any wearable sensors for data acquisition.
5. Studies that did not include DL methods for data analysis (e.g., feature extraction, classification, regression).
6. Studies that did not involve detection, monitoring, or prediction of response to treatment in PD.
7. Studies that did not use metrics that measure classification or regression performance.

3.4. Data extraction

Two authors (L.S. and L.B.) independently selected candidate studies by reviewing the title and abstract and repeated the process until they reached a consensus. The same procedure was performed for the selection based on the full-text evaluation. Finally, candidate studies that met the eligibility criteria were selected for inclusion in the review. The following information was included in the data extraction procedure:

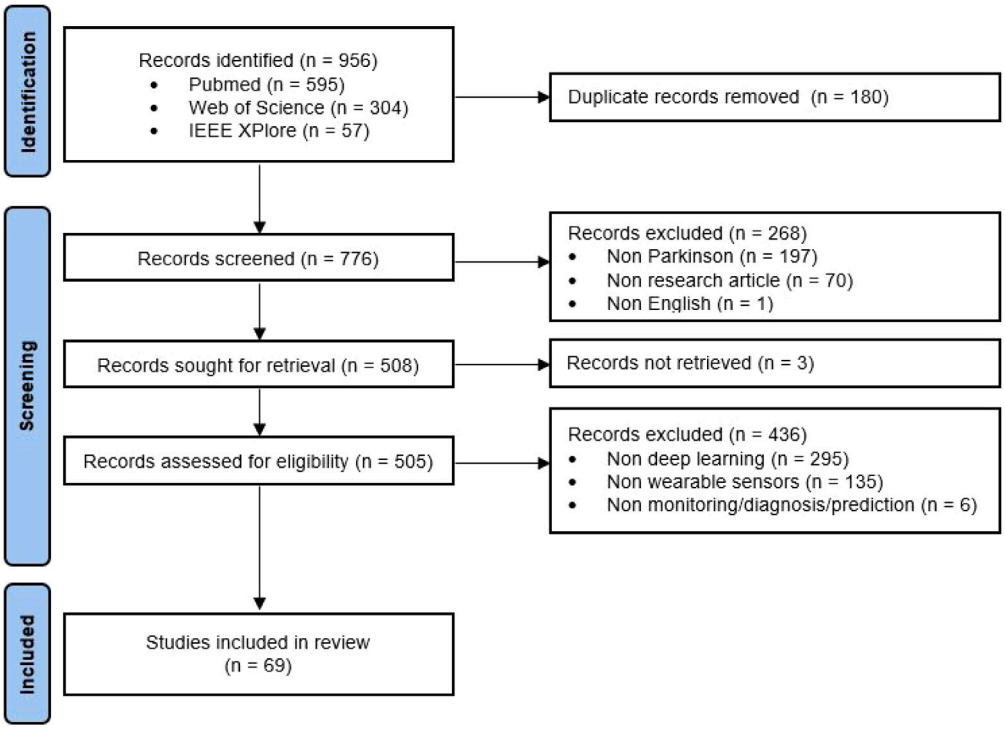

Fig. 1. PRISMA flow diagram of literature search and selection process showing the number of studies identified, screened, and included in the review.

(a) Study identification, including authors, title, and citation.
(b) Objective, including diagnosis, monitoring, or assessment of response to therapy.
(c) Characteristics of the data set, including the type of data set, sample size, and sociodemographic characteristics.
(d) Type, number, and location of the wearable sensors used for data acquisition.
(e) Motor or NMM being investigated.
(f) Information on the algorithmic approach, including preprocessing, segmentation, feature extraction, and feature selection.
(g) Deep learning architectures.
(h) Input data for classification or regression.
(i) Validation method.
(j) Performance evaluation metrics.

3.5. Quality report assessment

There are currently no quality scales for ML predictions. Therefore, the quality of the study was assessed using an adapted version of transparent reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) (Moons et al., 2015). For this task, 18 items (12 for methods and 6 for results) were assessed using an adapted TRIPOD checklist (Moons et al., 2015; Wang et al., 2020).

Most of the terms described in the TRIPOD documentation were used to assess the quality of the methods and results. In particular, three items were adapted (as reported in Wang et al. (2020)) to assess DL studies. In addition, two items related to blinded actions to assess outcomes and predictors (6b, 7b) were excluded because they did not apply to most of the studies examined. In particular, the reporting of the selection of hyperparameters was included in item 10b. Furthermore, item 15a was adapted to include the specification of links to the final model, coding of predictors or final parameters/coefficients, and with the architecture described in full in the article.

The adapted TRIPOD checklist is provided in Supplementary Table S1.

4. Results

This section reports the results of the systematic review process. Specifically, Section 4.1 describes the results of the literature search and the selection of studies, while Section 4.2 reports the quality assessment. The trend of publication, along with the main application areas, are reported in Section 4.3. (Section 4.4) presents the data source and sample size, while Section 4.5 summarizes the wearable devices and their positioning. In addition, input data type and data transformation methods (Section 4.6), DL methods (Section 4.7), model optimization procedures (Section 4.8), and performance evaluation procedures (Section 4.9) are reported and discussed. Furthermore, this section presents a detailed description of each motor symptom (Section 4.10) and NMM (Section 4.11) in which the use of wearables and DL was reported. Finally, a summary of the results is presented (Section 4.12).

4.1. Systematic review

Based on the search criteria, we retrieved 595 papers from PubMed, 57 from IEEE Xplore, and 304 from Web of Science, for a total of 956 publications. After removing duplicates (n = 180), we screened 776 publications for titles and abstracts, after which we excluded 268 based on the exclusion criteria. Of the remaining 508 records, 3 were excluded because full text was not available. We screened 505 full-text articles for eligibility and excluded 436 records according to the exclusion criteria. In the end, 69 research studies were included and analyzed. The PRISMA flowchart used for the literature search and selection is shown in Fig. 1.

4.2. Quality assessment

Fig. 2 shows the number of studies that correctly reported the corresponding TRIPOD assessment items. For each study, 18 items were assessed, with a maximum score of 18 (1 point for each item). The assessed items correspond to the reporting of methods and results.

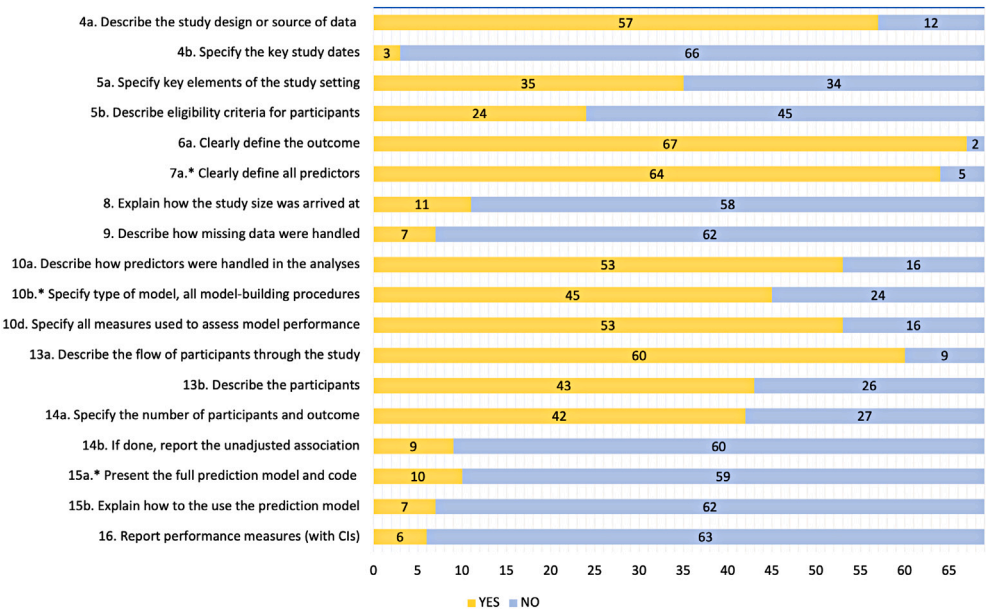

Fig. 2. Number of studies that reported the corresponding TRIPOD assessment items. YES=1, NO=0. The * indicates criteria adjusted for DL models.

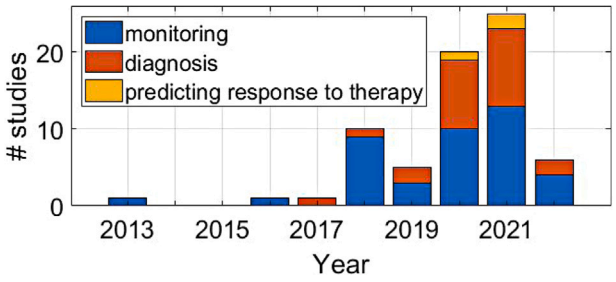

Fig. 3. Publication trend of collected research studies, along with their primary objective.

According to Fig. 2, none of the 69 studies received the maximum score (18 items). Thirty-two (46%) obtained scores between 10 and 15, and only two studies (3%) received the maximum score (15 items). The remaining thirty-seven studies (54%) received a score smaller than 10.

The design and source of the study data (4a), the study setting (5a), the definition of the outcome (6a), the definition of all predictors (7a), the description of how the predictors were handled (10a), the type of model (10b), the measures to evaluate the model (10d), the flow of participants (13a), the description of participants (13b) and the number of participants and the outcomes (14a) were adequately reported in most studies. However, key study dates (4b), participant eligibility criteria (5b), handling of missing data (9), unadjusted association between predictors and outcomes (14b), presentation of the full model (15a), explanation of how to use the predictive model (15b), and report of performance measures with confidence intervals were scarcely reported in most studies.

The use of TRIPOD guidelines indicates a lack of important information from the evaluated studies. Five items (4b, 9, 14b, 15b, 16) were poorly reported, some of which are related to the use of freely available databases. This situation makes it difficult to access specific information (e.g., key dates of studies or data handling) and hinders the accuracy of the information reported. In addition, the report of the unadjusted association between each candidate predictor and the outcome was often not reported. Furthermore, reporting of the hyperparameter selection process (included in item 10b) was often neglected, although the selection of an accurate set of hyperparameters can have a great impact on performance. In addition, the presentation of the complete model and the explanation of how to use it was rarely reported, a situation that hinders the reproduction of results and the comparison of different algorithmic approaches. On the contrary, clear definitions of predictors and outcomes were found, generally reporting the use of raw sensory signals and standardized rating scales.

4.3. Publication trend and objective

Fig. 3 shows the number of publications per year in the last decade, together with the application areas. Until 2017, very few studies (n = 3, 4.4%) used DL methods to analyze data from patients with PD (PwPD). Since 2018, a solid positive trend has been observed, with 15 articles (21.7%) published between 2018 and 2019. Most research papers (n = 51, 73.9%) were published in the last three years, with 45 studies (65.2%) published between 2020 and 2021. However, for 2022, there is not enough information to determine the trend, as this review considered studies published up to April 2022 (n = 6, 8.7%).

In the context of the evaluation of PD, previous studies (Deb, An, Bhat, Shill, & Ogras, 2022; Monje et al., 2019) classified the application areas of wearable technologies into the following primary areas: diagnosis, monitoring, and prediction of treatment response. Based on this classification, the application areas of research studies in which DL approaches were used in combination with wearables are shown in Fig. 3. Monitoring represents the main application (n = 41, 59%), with a positive publication trend since 2018. An even stronger trend is observed for computer-assisted diagnosis (n = 25, 37.2%), with most studies (n = 21) published in the last three years. Finally, some studies published in the last three years (n = 3, 4.4%) attempted to predict the response to treatment.

Among the signs of PD, motor symptoms have been investigated in 86% of the studies, while only 14% of the works addressed NMM (Fig. 4a). Fig. 4b shows the motor symptoms assessed by the studies under review. Gait impairment was found to be the most frequently investigated aspect (31%), followed by tremor and freezing of gait (FOG, 21% each), and bradykinesia (14%). Other symptoms addressed include dyskinesia (4%), fine motor impairment (4%), rigidity (1%), and balance (1%). Fig. 4c reports the NMM assessed by the collected research articles. Among them, speech impairment represents the most frequently investigated aspect (74%). Brain dysfunction was evaluated in 13% of the studies, while cognitive impairment and emotional expression dysfunction were addressed in 7% of the cases.

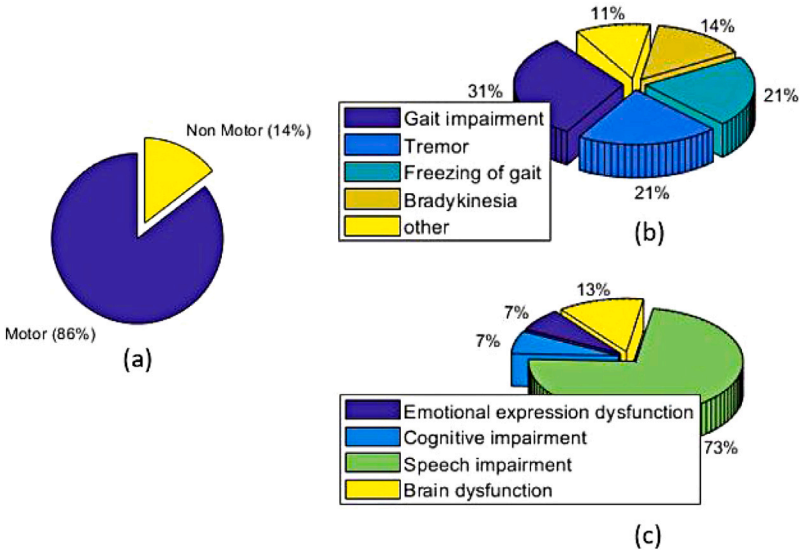

Fig. 4. Motor and non-motor symptoms investigated.

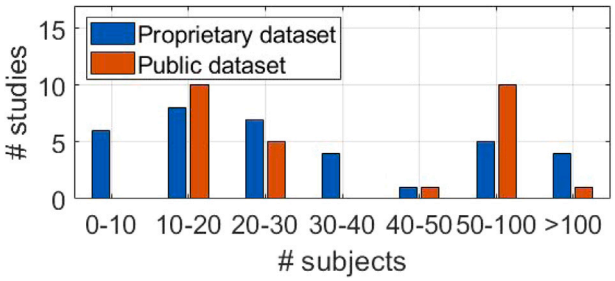

Fig. 5. Sample size in public and proprietary databases.

4.4. Data source and sample size

Research studies used private data sets (n = 37, 53.6%) more frequently than publicly available databases (n = 32, 46.4%). Fig. 5 shows the sample size for proprietary and public data sets. The number of PwPD enrolled ranges from 1 to 524 (mean: 55, median: 22) and from 10 to 188 (mean: 51, median: 21) for private and public databases, respectively. In general, most studies (92%) enrolled less than 50 PwPD, with only 5 studies (7.2%) using data collected from more than 100 patients.

The most frequently used publicly available databases included Physionet Gait in Parkinson's Disease (PhysioNet Gait) (n = 10, 31.2%) and Daphnet (n = 7, 21.9%). Physionet Gait in Neurodegenerative Disease Database (PhysioNet GNDD), REMPARK, and mPower data sets were used in two studies each (6.2%). The remaining data sets included Amigos, SEED-IV, CuPiD, Parkinson Speech data set with Multiple Types of Sound Recordings Data Set (Parkinson Speech), Parkinsons Data Set (Oxford), Parkinson Telemonitoring, Parkinson's disease classification, Parkinson's Speech, and Physical Activity Monitoring Data Set (PAMAP2), used each in a single study. Table 1 summarizes the most frequently used databases and their principal characteristics.

4.5. Wearable sensors

Different types of wearables were used for the analysis of motor symptoms and NMM in PD. Fig. 6a shows the sensing technology used in the studies reviewed. It should be noted that some studies used a combination of different sensors. Motion sensors were the most widely used; these sensors include accelerometers (37.7%), gyroscopes (22.6%), and magnetometers (6.6%), either alone or in combination with other sensors (e.g., EMG, force sensors), or integrated into devices such as smartphones, smartwatches, or smart bands. Force sensors were used in 14.2% cases, alone or in combination with an accelerometer and gyroscope. The use of EMG was relatively rare (0.9%), in a half cases in combination with accelerometers. Finally, the keypad and instrumented glove were used only once. The microphone was the sensor most commonly used to assess NMM (9.4%), representing one-third of all sensors used. Almost a fifth of the studies used the Emotiv headset (1.9%) or deep brain stimulation (1.9%) device to record cerebral activity.

Fig. 6b shows the position of the sensor on the human body. It should be noted that some studies used several sensors at different positions on the body. The size of the circles at each position is proportional to the frequency of occurrence.

The most common position for motion sensors (accelerometer, gyroscope, magnetometer) is the wrist (18.2%), followed by the lower leg (13.6%), waist (12.1%), upper leg (9.1%) and foot (6.1%). Other less frequent points are the hand, arm, pocket, chest, and finger. Force sensors were placed under the soles of the feet. The microphone was placed near the patients' mouths. Finally, the EEG headset was attached to the scalp of the subjects, and the EMG sensor to the patient's arm. Most of the studies (53%) used sensors embedded in commercially available devices, while the remaining studies used prototype sensors. Smartphone sensors (e.g., accelerometer, gyroscope, magnetometer, touchscreen, microphone) were used in 17.6% of the studies.

4.6. Feature engineering

Fig. 7 shows the type of input data used to train and evaluate the proposed DL architectures. On the one hand, given the capabilities of DL to process different data modalities, 38% of the studies used raw data to train and evaluate the proposed algorithmic approaches. On the other hand, 37% of the studies exclusively used hand-crafted features, including time domain features extracted from the raw time series and frequency domain features extracted after transforming the raw time series into the frequency domain.

Another data transformations methods reported in these studies include fast Fourier transform (FFT) (4%), continuous wavelet transform (4%), discrete Fourier transform (DFT) (1%), and Mel spectrogram (1%). The remaining studies used mixed data modalities, such as the

Table 1
List of the most frequently used databases. FOG: freezing of gait; IMU: inertial measurement unit; acc: accelerometer; gyro: gyroscope; mag: magnetometer; HC: Healthy control; ALS: amyotrophic lateral sclerosis; ADL: activities of daily living; HD: Huntington's disease; Parkinson's Speech: Parkinson Speech data set with Multiple Types of Sound Recordings Data Set; PhysioNet Gait: PhysioNet Gait in Parkinson's Disease; PhysioNet GNDD: PhysioNet Gait in Neurodegenerative Disease Database; n.r.: not reported.

Symptom	Database name	Database description	Device	Sensor	Number of sensors	Sensor location
Bradykinesia, gait, dysphonia	mPower (Sage Bionetworks, 2016)	Crowd-sourced data from 8003 subjects including memory, tapping, gait and voice-based tests	Smartphone	acc, touchscreen, microphone	3	Hand, mouth
Dysphonia	Parkinson's Disease Classification (Sakar et al., 2019)	Speech features extracted from samples of 188 PD and 64 HC	n.r.	Microphone	1	Mouth
Dysphonia	Parkinson Speech data set (Sakar et al., 2013)	20 HC and 20 PD pronouncing a list of numbers, words, and sustained vowels	Consumer microphone	Microphone	1	Mouth
Dysphonia	Parkinsons Telemonitoring (Tsanas, Little, McSharry, & Ramig, 2009)	Speech features extracted from speech of 42 early-stage PD subjects in a six-month test	Smartphone	Microphone	3	Mouth
Dysphonia	Parkinsons Data Set (Oxford) (Little, Mcsharry, Roberts, Costello, & Moroz, 2007)	Speech features from 31 people (23 PD)	Head-mounted microphone	Microphone	1	Mouth
FOG	Daphnet (Bachlin et al., 2009)	Gait and FOG measurements of 10 PD subjects	Prototype IMU	acc, gyro, mag	3	Waist, upper-leg, lower-leg
FOG	CuPiD (Mazilu et al., 2013)	Gait and FOG measurements of 18 PD subjects	Prototype IMU	acc, gyro	1	Wrist
FOG	REMPARK (Rodríguez-Martín et al., 2017)	Gait and FOG measurements of 21 PD subjects performing scripted ADL	Prototype IMU	acc, gyro	1	Wrist
Gait impairment	PhysioNet Gait (Goldberger et al., 2000)	Gait measurements of 93 PD and 73 HC	Prototype	Force	16	Feet
Gait impairment	PhysioNet GNDD (Hausdorff et al., 2000)	Gait database with 13 ALS, 20 HD, 15 PD, and 16 HC	Prototype	Force	16	Feet

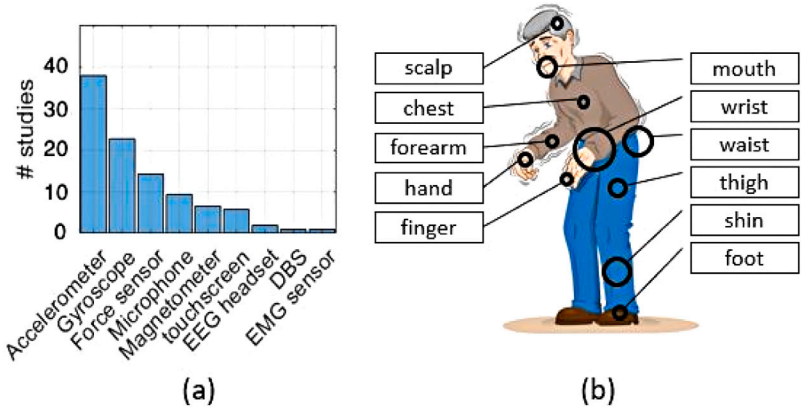

Fig. 6. Type and body-location of sensing technology. EEG: electroencephalography; EMG: electromyography; DBS: deep brain stimulation.

combination of raw signals with hand-crafted features (9%) or the application of principal component analysis (PCA) to extracted features (4%).

4.7. Deep learning methods

Fig. 8 shows the DL models used in the studies investigated. CNNs represent more than half (52%) of the implemented DL algorithms. RNN and multi-layer perceptron (MLP) were used in 24% and 14% of the studies, respectively, while deep autoencoders (DAE) were used in 7% of the studies. The remaining studies reported miscellaneous DL architectures.

More in detail, most of the proposed DL architectures to assess motor symptoms and NMM used CNN-based architectures. These architectures used CNNs as discriminative blocks to learn features directly from raw data without the need for a manual feature extraction process. Furthermore, these CNN blocks included pooling operations (e.g., maximum, average, or global pooling) commonly placed after the convolution to reduce the size of feature maps. Finally, fully connected

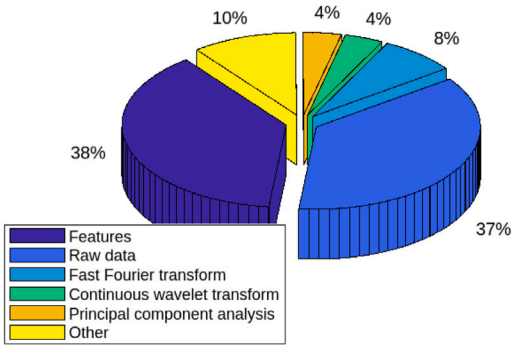

Fig. 7. Pie chart of data used to feed the deep learning algorithms.

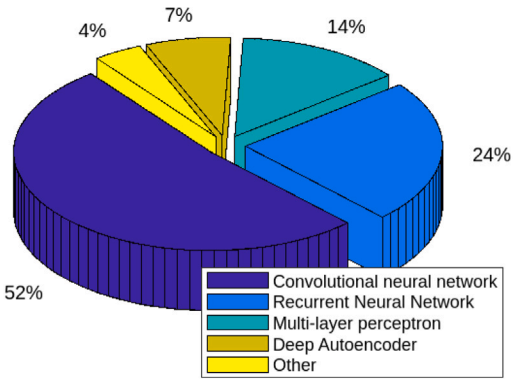

Fig. 8. Pie chart of the deep learning algorithms.

layers blocks were placed at the top of CNN blocks to perform classification or regression tasks. In the reviewed studies, standard CNN-based architectures followed a pyramid-inverted scheme, where the size of the input data was gradually reduced by means of convolution operations, pooling operations, and fully connected blocks to produce a result corresponding to the expected prediction (classification or regression).

Meanwhile, the RNN-based architectures used LSTM layers (22%) for sequential modeling followed by fully connected blocks to produce a prediction output. Other recurrent processing layers used in the studies under review include bidirectional LSTM (6%), GRU (2%), and convolutional recurrent neural network (CRNN) (1%). Furthermore, only 3% of the studies reported the use of the attention mechanism (Niu, Zhong, & Yu, 2021) in combination with RNN.

The use of DAE was reported in 7% of the studies. Most of these studies used DAE to learn efficient data representations from raw signals in an unsupervised fashion. The DAE were commonly implemented using a combination of CNN and fully connected layers. In general, CNNs (4%) and fully connected layers (3%) were used to implement encoder and decoder blocks, while a single fully connected layer was used as latent space for dimensional reduction.

Other studies reported the use of multilayer perceptron architectures composed of fully connected layers, generally used to process hand-crafted features. Furthermore, miscellaneous DL architectures including sequence-optimized modular neural networks, dynamic neural networks, probabilistic neural networks and graph neural networks were proposed and evaluated.

4.8. Hyperparameter optimization

The performance of ML and DL models depends on their hyperparameters. During DNN training, the weights and biases of the network layers are updated; however, this process is difficult due to the number of parameters to be adjusted in each layer (Glorot & Bengio, 2010).

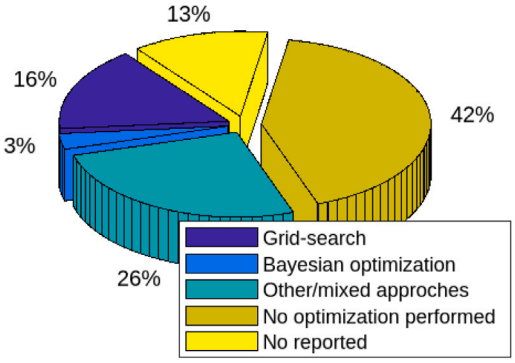

Fig. 9. Pie chart of the hyperparameter optimization approaches.

In the DL context, hyperparameters refer to parameters that cannot be updated during the training of a model (i.e. the number of hidden layers, layer type, layer parameters, activation function, learning rate, batch size, or optimizer algorithm). Accurate selection (optimization) of these hyperparameters has a great influence on model performance and controls the training process in terms of computational processing.

Fig. 9 shows the hyperparameter optimization approaches reported in the studies under investigation. In this review, 45% of studies reported the implementation of hyperparameter optimization, 13% did not perform hyperparameter optimization, and the remaining 42% did not report the hyperparameter optimization process in their studies. Among the studies that reported the hyperparameter optimization process, 16% used the grid search method, 3% employed Bayesian optimization, 1% used the Hyperband method, 1% the Broyden–Fletcher–Goldfarb–Shanno (BFGS), and the remaining 24% reported the use of different techniques including manual search, Tubu search bias value optimization or mixed approaches.

4.9. Performance evaluation

Classification tasks were addressed in 90% of the studies, while regression tasks were performed in only 10% of the studies. Fig. 10a reports the main classification metrics calculated in the studies reviewed. Accuracy was the most frequent metric (68.8%), followed by sensitivity (57.4%) and specificity (50.8%). The area under the curve (AUC) was reported in 32.8% of the cases, F-score in 27.9%, and precision in 22.9%. Other classification metrics include the equal error rate (EER, 2 studies) and Matthew's correlation coefficient (MCC, 4 studies).

Fig. 10b shows the most employed regression metrics. The Pearson's correlation coefficient (r) was reported in all studies. More than a half of the studies (57.1%) also reported the mean absolute error (MAE), while root mean square error (RMSE), and mean square error (MSE) were reported in 28.6% and 14.3% of the studies, respectively. Furthermore, two studies assessed inter-rater reliability in evaluating specific motor symptoms. To this end, the intra-class correlation coefficient (ICC) or the Cohen's kappa (k) coefficient were reported.

4.10. Motor symptoms

Several studies employed DL approaches to assess motor manifestations and cardinal symptoms of PD. This section describes technological solutions based on wearable sensors and DL methods to assess the different motor aspects of PD, including FOG (Section 4.10.1), gait impairment (Section 4.10.2), tremor (Section 4.10.3), bradykinesia (Section 4.10.4), and dyskinesia (Section 4.10.5). A summary of each of these symptoms is presented in the following subsections, where the data set, signal processing techniques, DL methods, validation approaches, and results are reported and briefly discussed.

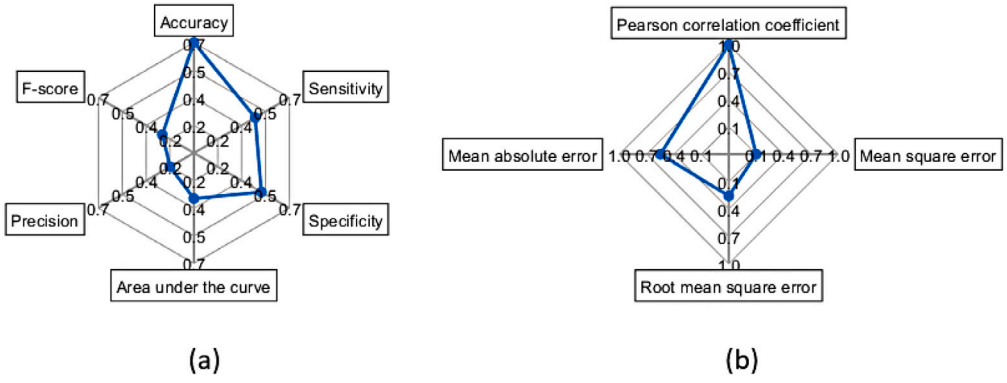

Fig. 10. Radar plot of the classification (a) and regression (b) metrics used for performance evaluation.

4.10.1. Freezing of gait

Fifteen studies used wearables and DL methods for FOG detection. Information on the sample, sensor technology, methods, and results are summarized in Table 2 and discussed in detail below.

Data. The number of participants enrolled ranged from 7 (Naghavi & Wade, 2022; O'Day et al., 2022) to 63 (Shi et al., 2022). In more detail, seven studies (Ashfaque Mostafa, Soltaninejad, McIsaac, & Cheng, 2021; Ashour, El-Attar, Dey, El-Kader, & Abd El-Naby, 2020; Esfahani, Dyka, Ortmann, & Langendörfer, 2021; Li et al., 2020; Mohammadian Rad, Van Laarhoven, Furlanello, & Marchiori, 2018; Noor, Nazir, Wahab, & Ling, 2021; San-Segundo, Navarro-Hellín, Torres-Sánchez, Hodgins, & De la Torre, 2019) used the Daphnet (Bachlin et al., 2009) data set, which includes acceleration data recorded from 10 PwPD using three inertial sensors placed on the lower back, thigh, and ankle. Participants were asked to perform three different types of walking tasks that simulate activities of daily living (ADL). A total of 237 episodes of FOG were recorded during the experiments. Two studies (Camps et al., 2018; Sigcha et al., 2020) used the REMPARK data set (Rodríguez-Martín et al., 2017), comprising data from 21 PwPD recorded with a single inertial measurement unit (IMU) mounted at the waist. The subjects performed various programmed activities, simulating those of daily life. A total of 1058 FOG episodes were collected during the recordings. The CUPID data set (Mazilu et al., 2013) was used in Kim et al. (2018). It includes accelerometer and gyroscope recordings from an IMU placed on the wrists of 11 patients while they performed walking, turning, and obstacle negotiation tasks. The data collected included a total of 184 episodes of FOG. In Shi et al. (2022), 63 PwPD were asked to perform a 7-m timed-up-and-go (TUG) test and a free-walking task while wearing two IMUs on their ankles. A total of 486 FOG episodes were recorded. In Naghavi and Wade (2022), two IMUs were applied to the ankles of 7 PwPD while they walked in a narrow corridor. During the experiments, 154 episodes of FOG were recorded. In Shalin, Pardoel, Lemaire, Nantel, and Kofman (2021), plantar pressure data from 11 PwPD were recorded while walking a predefined path that included entering a narrow corridor, turning at different angles, and voluntary stops. In total, 362 FOG events were registered. In Kim et al. (2018), a smartphone was placed in different positions (e.g., pocket, waist, and lower limb) and used to collect inertial data from 32 PwPD. Participants were asked to perform various walking tasks, including a 3-m TUG, turning around, and opening and entering a door. The experiments resulted in the collection of 110 episodes of FOG. Finally, in O'Day et al. (2022), six IMUs were attached to the top of both feet, the lateral side of both shins, and the lumbar (L5) and thoracic regions. Inertial data were recorded while participants performed several walking trials consisting of two ellipses and two figures of eight around high barriers. A total of 211 FOG episodes were recorded.

Pre-processing. Low-pass (Camps et al., 2018), high-pass (San-Segundo et al., 2019), and band-pass (Naghavi & Wade, 2022; Sigcha et al., 2020) filters were used to remove low-frequency trends and high-frequency noise. Other processing techniques included data normalization (Noor et al., 2021; O'Day et al., 2022; Shi et al., 2022), outlier filtering (Li et al., 2020; Noor et al., 2021), data balancing (Ashfaque Mostafa et al., 2021; Li et al., 2020; Shalin et al., 2021), and data augmentation (Camps et al., 2018; O'Day et al., 2022). Thirteen studies used fixed-length windows for the segmentation of inertial data into temporal frames, with time windows ranging from 1 s to 4 s. Specifically, most studies used window sizes between 2 s and 3 s (Ashfaque Mostafa et al., 2021; Bikias et al., 2021; Camps et al., 2018; Kim et al., 2018; Naghavi & Wade, 2022; Noor et al., 2021; O'Day et al., 2022), followed by longer windows (up to 4 s) (Li et al., 2020; San-Segundo et al., 2019; Shi et al., 2022; Sigcha et al., 2020), while windows shorter than 1 s were sparsely used (Esfahani et al., 2021; Mohammadian Rad et al., 2018). An overlap between 10% and 75% was used in eight studies (Bikias et al., 2021; Camps et al., 2018; Li et al., 2020; Mohammadian Rad et al., 2018; Naghavi & Wade, 2022; San-Segundo et al., 2019; Shi et al., 2022; Sigcha et al., 2020), while three studies did not use any overlap (Ashfaque Mostafa et al., 2021; Esfahani et al., 2021; Noor et al., 2021). Finally, single-sample classification was exploited in Shalin et al. (2021).

Deep learning models. Eight studies (Ashfaque Mostafa et al., 2021; Bikias et al., 2021; Camps et al., 2018; Kim et al., 2018; Naghavi & Wade, 2022; O'Day et al., 2022; San-Segundo et al., 2019; Shi et al., 2022) used CNN fed with raw data (Bikias et al., 2021; Naghavi & Wade, 2022; O'Day et al., 2022; San-Segundo et al., 2019), CWT (Shi et al., 2022), DWT (Ashfaque Mostafa et al., 2021), and FFT (Camps et al., 2018; Kim et al., 2018). Three studies (Ashour et al., 2020; Esfahani et al., 2021; Shalin et al., 2021) used LSTM fed with raw data (Ashour et al., 2020) and kinematic features (Shalin et al., 2021); two articles (Li et al., 2020; Sigcha et al., 2020) exploited a combination of both models, using raw data (Li et al., 2020) and the FFT (Sigcha et al., 2020) as input; finally, two studies used a convolutional denoising autoencoder (CDA) (Mohammadian Rad et al., 2018; Noor et al., 2021) fed with raw data.

Validation procedure. Six studies (Bikias et al., 2021; Li et al., 2020; Naghavi & Wade, 2022; San-Segundo et al., 2019; Shalin et al., 2021; Sigcha et al., 2020) used a leave-one-subject-out (LOSO) validation to evaluate the performance of the proposed system. Five studies (Ashour et al., 2020; Camps et al., 2018; Esfahani et al., 2021; Noor et al., 2021; Shi et al., 2022) used hold-out validation. Of these, four studies (Ashour et al., 2020; Camps et al., 2018; Noor et al., 2021; Shi et al., 2022) used an S-I approach, while only one study (Esfahani et al., 2021) used an S-D approach. Finally, two studies (Ashfaque Mostafa et al., 2021; Kim et al., 2018) used k-fold CV. In particular, 5- and 10-fold CV was used in Ashfaque Mostafa et al. (2021) and Kim et al. (2018), respectively. Both studies used an RS strategy to split the data set before validation, without guaranteeing the independence of the subjects in the training and test subsets.

Table 2
List of studies that used wearable sensors and DL methods for freezing of gait detection (FOG). ACC: accuracy; Sens: sensitivity; Spec: specificity; AUC: area under the curve; acc: accelerometer; gyro: gyroscope; CWT: continuous wavelet transform; CNN: convolutional neural network; LSTM: Long-short term memory; ANN: artificial neural; LOSO: leave-one-subject-out; CDA: convolutional denoising autoencoder; DWT: discrete wavelet transform; S-I: subject independent; S-D: subject dependent; RS: random shuffle.

Study	# PD subjects	Sensor type (position)	Model (Input)	Validation	Performance
Shi et al. (2022)	63	acc, gyro (ankles)	CNN (CWT)	Hold-out (S-I)	Sens 0.878 Spec 0.864 ACC 0.871
Esfahani et al. (2021)	10	acc (ankle)	LSTM (raw data)	Hold-out (S-D)	Sens 0.926 Spec 0.956 AUC 0.976
Naghavi and Wade (2022)	7	acc, gyro (ankles)	CNN (raw data)	LOSO (S-I)	Sens 0.630 Spec 0.986
Noor et al. (2021)	10	acc (thigh)	CDA (raw data)	Hold-out (S-I)	Sens 0.909 Spec 0.670 ACC 0.790
Ashour et al. (2020)	10	acc (back, thigh, ankle)	LSTM (raw data)	Hold-out (S-I)	ACC 0.834
Shalin et al. (2021)	11	force (feet)	LSTM (16 features)	LOSO (S-I)	Sens 0.821 Spec 0.895 F-score 0.350
Li et al. (2020)	10	acc (back, thigh, ankle)	CNN+LSTM (raw data)	LOSO (S-I)	Sens 0.875 Spec 0.923 ACC 0.920
Sigcha et al. (2020)	21	acc (waist)	CNN+LSTM (FFT)	LOSO (S-I)	Sens 0.871 Spec 0.871 AUC 0.939
Ashfaque Mostafa et al. (2021)	10	acc (back)	CNN (DWT)	k-fold (RS)	Sens 0.946 Spec 0.952 ACC 0.946
San-Segundo et al. (2019)	10	acc (back, thigh, ankle)	CNN (raw data)	LOSO (S-I)	Sens 0.950 Spec 0.725 AUC 0.930
Bikias, Iakovakis, Hadjidimitriou, Charisis, and Hadjileontiadis (2021)	11	acc, gyro (wrists)	CNN (raw data)	LOSO (S-I)	Sens 0.830 Spec 0.880
Kim et al. (2018)	32	acc, gyro (pocket)	CNN (FFT)	k-fold (RS)	Sens 0.938 Spec 0.901 F-score 0.920
Camps et al. (2018)	21	acc, gyro (waist)	CNN (FFT)	Hold-out (S-I)	Sens 0.919 Spec 0.895 ACC 0.890
O'Day et al. (2022)	7	acc, gyro (back, ankles)	CNN (raw data)	LOSO (S-I)	AUC 0.830
Mohammadian Rad et al. (2018)	10	acc (back, thigh, ankle)	CDA (raw data)	Hold-out (S-D)	AUC 0.770

Results. Most studies reported results in terms of sensitivity and specificity, while two studies (O'Day et al., 2022; Peraza et al., 2021) provided only the area under the curve (AUC) and one (Ashour et al., 2020) the accuracy alone. Sensitivity and specificity ranged from 0.63–0.95 and 0.67–0.99, respectively. The geometric mean between sensitivity and specificity (GM) ranged from 0.78 to 0.95. Only three studies (Ashfaque Mostafa et al., 2021; Camps et al., 2018; Esfahani et al., 2021; Kim et al., 2018) reported a GM greater than 0.90. However, when considering the type of validation, in Ashfaque Mostafa et al. (2021) and Kim et al. (2018) RS k-fold CV was used, thus reducing the reliability of the results; in Esfahani et al. (2021) a subject-specific (S-D) algorithm was implemented so that a direct comparison with similar studies is not possible; in Camps et al. (2018), a hold-out test was used, in which the classification model was tested on 4 new PwPD. In terms of DL algorithms, CDA performed poorly (Mohammadian Rad et al., 2018; Noor et al., 2021); CNN performed better than LSTM, on average; finally, the combination of CNN and LSTM produced sensitivity and specificity above 0.87 in LOSO CV (Li et al., 2020; Sigcha et al., 2020).

4.10.2. Gait impairment

Twenty-two studies used wearables and DL methods to assess gait impairment. Eighteen studies specifically focused on gait analysis, while the remaining four analyzed gait impairment in addition to bradykinesia, dysphonia, and cognitive impairment (Prince, Andreotti, & De Vos, 2019); bradykinesia and tremor (Stamate et al., 2018); balance (Moon et al., 2020) and bradykinesia; and dysphonia (Deng et al., 2022). The information on the samples, sensor technology, methods, and results are summarized in Table 3 and discussed in detail below.

Data. The number of PwPD enrolled ranged from 12 (Stamate et al., 2018) to 645 (Deng et al., 2022). Most of the articles included healthy controls (HC) in their studies, except two articles (Moon et al., 2020; Stamate et al., 2018). In more detail, ten studies (Alharthi, Casson, & Ozanyan, 2021; Balaji, Brindha, Vinodh Kumar, & Vikrama, 2021; El Maachi, Bilodeau, & Bouachir, 2020; Jane, Nehemiah, & Arputharaj, 2016; Liu, Li, Liu, Du, & Zo, 2021; Oğul & Özdemir, 2022; Pham, 2021; Setiawan & Lin, 2021; Xia, Yao, Ye, & Cheng, 2020; Zhao, Qi, Li, Dong, & Yu, 2018) used the PhysioNet Gait database (Goldberger

et al., 2000), two studies (Berke Erdas, Sumer, & Kibaroglu, 2022; Lin, Wen, & Setiawan, 2020) used the PhysioNet GNDD (Hausdorff et al., 2000), two studies (Deng et al., 2022; Prince et al., 2019) used data collected through the mPower application, seven studies (Chen, Fan, Li, Zou, & Huang, 2021; Fernandes et al., 2021; Lin et al., 2022; Moon et al., 2020; Peraza et al., 2021; Stamate et al., 2018; Steinmetzer, Maasch, Bönninger, & Travieso, 2019) used proprietary databases, and a single study (Zhao et al., 2022) used a combination of three data sources (e.g., Daphnet Bachlin et al., 2009, PhysioNet Goldberger et al., 2000, and PhysioNet GNDD Hausdorff et al., 2000). The PhysioNet Gait data set includes gait data from 93 patients with idiopathic PD and 73 HC. Eight Vertical Ground Reaction Force (VGRF) sensors were placed below the foot of each patient to record data at a sampling rate of 100 Hz. As part of the data collection protocol, subjects were instructed to walk back and forth to the corridor at a normal pace for approximately two minutes. The PhysioNet GNDD consists of gait data from 64 subjects (13 with amyotrophic lateral sclerosis, 20 with Huntington's, 15 with PD, and 16 HC). Data were collected using force-sensitive resistors placed under the feet. In Prince et al. (2019), data collected from 377 PwPD and 247 HC were used for gait assessment. Data were recorded via the mPower mobile application while the subjects performed various activities (e.g., walking back and forth, tapping their feet, and performing memory and vocal tasks). In Deng et al. (2022), data from the DREAM Parkinson's disease challenge (mPower) (Sieberts et al., 2021) were used. Data used for gait assessment included 645 PwPD and 2084 HC performing 30-step walking trials. With regard to proprietary databases, in Lin et al. (2022) data from 32 PwPD and 16 HC were recorded using an IMU while the participants walked back and forth for 10 m. In Steinmetzer et al. (2019), two smart bracelets were used to collect arm swing measurements of 15 PwPD and 24 HC performing the TUG test. In Chen et al. (2021), data were collected from 50 PwPD and 50 HC walking in a straight line for 50 meters with a prototype IMU attached to their foot. In Stamate et al. (2018), data were recorded from 12 PwPD while they performed MDS-UPDRS tasks; this protocol included walking in a straight line for 5 m, turning, and returning to the starting point while carrying a smartphone in their belt or trouser pocket. In Moon et al. (2020), 524 PwPD were asked to wear six IMUs while performing walking and standing tests. In Peraza et al. (2021), 6 PwPD and 6 HC were equipped with four IMUs on the wrist, waist, and legs, and asked to perform different walking tasks (slow, normal, fast walking, and TUG test). In Fernandes et al. (2021), gait recordings of 14 subjects with vascular PD, 15 with idiopathic PD, and 34 HC were recorded in both ON and OFF states while walking for 60 m. For this task, two wearable motion sensors were attached to the back of each shoe with two elastic bands. In Zhao et al. (2022), three data sources were used, including PhysioNet Gait in Parkinson's Disease (Goldberger et al., 2000), PhysioNet Gait in NDD (Hausdorff et al., 2000), and Daphnet (Bachlin et al., 2009). For the latter, acceleration measurements of PwPD performing walking tasks (e.g., walking in a straight line, turning, and performing ADL tasks) were recorded using three sensors (triaxial accelerometers) attached to the participants' hips and legs.

Pre-processing. Band-pass filtering (Peraza et al., 2021) was used to process inertial data. Other processing techniques applied to gait assessment include data normalization (Alharthi et al., 2021; Chen et al., 2021; Prince et al., 2019; Setiawan & Lin, 2021; Zhao et al., 2018), zero-padding (Xia et al., 2020), QR code transformation (Berke Erdas et al., 2022), and data balancing (Pham, 2021). Regarding signal segmentation, 15 studies used fixed-length windows between 0.8 s (Lin et al., 2022) and 12 s (Jane et al., 2016). Specifically, most studies used windows between 0.8 s and 5 s (Alharthi et al., 2021; Balaji et al., 2021; Chen et al., 2021; El Maachi et al., 2020; Lin et al., 2022; Liu et al., 2021; Peraza et al., 2021; Stamate et al., 2018; Steinmetzer et al., 2019; Xia et al., 2020; Zhao et al., 2022, 2018), while windows of 12 s (Jane et al., 2016), 10 s (Lin et al., 2020; Setiawan & Lin, 2021) or longer (10 s, 15 s, and 30 s) (Setiawan & Lin, 2021) were used. In addition, an overlap of 50% was used (Chen et al., 2021; El Maachi et al., 2020; Stamate et al., 2018), 60% in Balaji et al. (2021), and 70% in Lin et al. (2020). The remaining studies (Berke Erdas et al., 2022; Deng et al., 2022; Fernandes et al., 2021; Moon et al., 2020; Oğul & Özdemir, 2022; Pham, 2021; Prince et al., 2019; Setiawan & Lin, 2021) did not use any sliding window strategy. Instead, they used data from the entire walking session to evaluate the algorithms.

Deep learning models. Thirteen studies (Alharthi et al., 2021; Berke Erdas et al., 2022; Chen et al., 2021; Deng et al., 2022; El Maachi et al., 2020; Fernandes et al., 2021; Lin et al., 2022, 2020; Peraza et al., 2021; Prince et al., 2019; Setiawan & Lin, 2021; Stamate et al., 2018; Steinmetzer et al., 2019) used CNN with raw data (Alharthi et al., 2021; Chen et al., 2021; Deng et al., 2022; El Maachi et al., 2020; Lin et al., 2022, 2020; Stamate et al., 2018), CWT (Setiawan & Lin, 2021; Steinmetzer et al., 2019), Mel's spectrogram (Peraza et al., 2021), and specific features (Berke Erdas et al., 2022; Fernandes et al., 2021; Prince et al., 2019). Furthermore, seven studies (Balaji et al., 2021; Liu et al., 2021; Oğul & Özdemir, 2022; Pham, 2021; Xia et al., 2020; Zhao et al., 2022, 2018) used RNN architectures. Among this group, two studies (Balaji et al., 2021; Xia et al., 2020) used LSTM networks fed with raw data, and one study used features (Pham, 2021). Furthermore, in Oğul and Özdemir (2022) the authors introduce a Siamese recurrent network with attention (RSRNA) fed with raw data. Another paper introduced a recurrent model named correlative memory neural network (CorrMNN) fed with features (Zhao et al., 2022). Furthermore, the introduction of a dual-branch CNN with BiLSTM (Liu et al., 2021) and the combination of CNN with LSTM (Zhao et al., 2018) fed with raw signals have been reported. Finally, two studies used ANN fed with features (Moon et al., 2020) or raw signals (Jane et al., 2016).

Validation. Several validation methods were used to estimate the performance of the algorithms, including k-fold (Berke Erdas et al., 2022; Chen et al., 2021; Deng et al., 2022; El Maachi et al., 2020; Fernandes et al., 2021; Jane et al., 2016; Lin et al., 2022; Moon et al., 2020; Oğul & Özdemir, 2022; Pham, 2021; Prince et al., 2019; Setiawan & Lin, 2021; Steinmetzer et al., 2019; Xia et al., 2020; Zhao et al., 2018), hold-out (Alharthi et al., 2021; Balaji et al., 2021; Liu et al., 2021; Zhao et al., 2022), and LOSO CV (Lin et al., 2020; Peraza et al., 2021; Stamate et al., 2018). In more detail, most studies used k-fold strategies, including stratified k-fold (Fernandes et al., 2021; Prince et al., 2019; Xia et al., 2020), k-fold with random shuffling (Berke Erdas et al., 2022; Chen et al., 2021; Jane et al., 2016; Lin et al., 2022; Moon et al., 2020; Pham, 2021; Setiawan & Lin, 2021; Steinmetzer et al., 2019; Zhao et al., 2018), and k-fold with subject-independent data split (Deng et al., 2022; El Maachi et al., 2020; Oğul & Özdemir, 2022). Specifically, 5-fold strategies were used in Chen et al. (2021), Deng et al. (2022), Fernandes et al. (2021), Lin et al. (2022) and Xia et al. (2020), 10-fold strategies in Berke Erdas et al. (2022), El Maachi et al. (2020), Jane et al. (2016), Oğul and Özdemir (2022), Pham (2021), Prince et al. (2019), Setiawan and Lin (2021), Zhao et al. (2018), and 3-fold strategies in Moon et al. (2020) and Steinmetzer et al. (2019). Finally, four studies (Alharthi et al., 2021; Liu et al., 2021; Zhao et al., 2022, 2018) used a hold-out split approach, randomly dividing the data set into training sets and evaluation sets.

Results. Regression models were used in two studies (Berke Erdas et al., 2022; Peraza et al., 2021) to predict the severity of gait impairment (Berke Erdas et al., 2022) and to extract gait parameters (e.g., step length, stride speed, stride length, and stride velocity) (Peraza et al., 2021), while the remaining works aimed to classify the presence or severity of gait impairment. In Berke Erdas et al. (2022), Pearson's correlation coefficient r of 0.79 was obtained when comparing the prediction of the model with the severity of gait impairment, while in Peraza et al. (2021) r coefficient ranged from 0.753 to 0.892 were reported. For the classification task, accuracy was the most reported metric. According to Table 3, an accuracy of up to 1 was reported (Pham, 2021) using a k-fold strategy (random shuffling),

Table 3
List of studies using wearable sensors and DL methods for the assessment of gait impairment. ACC: accuracy; Sens: sensitivity; Spec: specificity; AUC: area under the curve; RMSE: root mean square error; MSE: mean square error; MAE: mean absolute error; acc: accelerometer; gyro: gyroscope; mag: magnetometer; CNN: convolutional neural network; RNN: recurrent neural network; LSTM: long short-term memory; BiLSTM: Bi-directional long short-term memory; ANN: artificial neural network; CWT: continuous wavelet transform; LOSO: leave-one-subject-out; S-I: subject independent; RS: random shuffle.

Study	# PD subjects (# Controls)	Sensor type (position)	Model (Input)	Validation type	Performance
Xia et al. (2020)	93 (73)	force (feet)	LSTM with Attention (raw data)	k-fold (stratified)	ACC 0.990
Alharthi et al. (2021)	93 (73)	force (feet)	CNN (raw data)	Hold-out (S-I)	ACC 0.955 F-score 0.960
Lin et al. (2022)	32 (16)	acc, gyro, mag (waist, leg, wrist)	CNN (raw data)	k-fold (RS)	ACC 0.997 Sens 0.995 Spec 0.900
Prince et al. (2019)	377 (247)	acc, gyro (waist, hand)	CNN (features)	k-fold (stratified)	ACC 0.774 F-1 score 0.830
Oğul and Özdemir (2022)	93 (73)	force (feet)	RNN with Attention (raw data)	k-fold (S-I)	ACC 0.810 AUC 0.878
Zhao et al. (2022)	108 (89)	force, acc, image (feet, waist, leg)	RNN (features)	Hold-out (S-I)	ACC 0.989
Liu et al. (2021)	93 (73)	force (feet)	CNN,BiLSTM (raw data)	Hold-out (S-I)	ACC 0.992 Sens 1.0 Spec 0.980
Zhao et al. (2018)	93 (73)	force (feet)	CNN, LSTM (raw data)	k-fold (RS)	ACC 0.988
Balaji et al. (2021)	93 (73)	force (feet)	LSTM (raw data)	Hold-out (S-I)	ACC 0.986 Sens 0.982 Spec 0.991
El Maachi et al. (2020)	93 (73)	force (feet)	CNN (raw data)	k-fold (S-I)	ACC 0.853 F-score 0.850
Steinmetzer et al. (2019)	15 (24)	acc, gyro, mag (wrist)	CNN (CWT)	k-fold (RS)	ACC 0.933 Sens 0.934 Spec 0.932 F-score 0.930
Berke Erdas et al. (2022)	15 (16)	force (feet)	CNN (features)	k-fold (RS)	RMSE 0.92 MSE 0.850 MAE 0.710 r 0.79
Chen et al. (2021)	50 (50)	acc, gyro (feet)	CNN (raw data)	k-fold (RS)	ACC 0.914 AUC 0.93 Sens 0.901 Spec 0.917
Stamate et al. (2018)	12	acc (pocket, leg, hand, finger)	Recurrent-CNN (raw data)	LOSO (S-I)	ACC 0.780 AUC 0.870 F-score 0.820
Moon et al. (2020)	524	acc, gyro, mag (wrist, feet, chest, waist)	ANN (features)	k-fold (stratified)	ACC 0.89 Sens 0.61 F-score 0.61
Jane et al. (2016)	93 (73)	force (feet)	ANN (raw data)	k-fold (RS)	ACC 0.922 RMSE 0.615
Peraza et al. (2021)	6 (6)	acc (wrist, waist, leg)	CNN (Mel spectrogram)	LOSO (S-I)	MAE 22.82 r>0.75
Fernandes et al. (2021)	29 (34)	acc, gyro, force (feet)	CNN (features)	k-fold (stratified)	ACC 0.860 Sens 0.800 Spec 0.900
Deng et al. (2022)	645 (2084)	acc (hand)	CNN (raw data)	k-fold (S-I)	AUC 0.898
Pham (2021)	93 (73)	force (feet)	LSTM (features)	k-fold (RS)	AUC 1.0 Sens 1.0 Spec 1.0
Lin et al. (2020)	15 (16)	force (feet)	CNN (raw data)	LOSO (S-I)	ACC 0.967 AUC 0.960 Sens 0.932 Spec 0.978
Setiawan and Lin (2021)	93 (73)	force (feet)	CNN (CWT)	k-fold (RS)	ACC 0.983 AUC 0.980 Sens 0.977 Spec 0.988

while the Hold-out (subject-independent data shuffling) and LOSO approaches provided an accuracy of 0.992 (Liu et al., 2021) and 0.967 (Lin et al., 2020), respectively. Higher accuracy values were reported using force sensors from the PhysioNet Gait in Parkinson's Disease database (0.810 (Oğul & Özdemir, 2022) to 1 (Liu et al., 2021)) and PhysioNet Gait in NDD (0.967 (Oğul & Özdemir, 2022) to 0.989 (Liu et al., 2021)). However, the conservative results reported in a subject-independent k-fold validation (El Maachi et al., 2020; Oğul & Özdemir, 2022) provided an accuracy of up to 0.853 (El Maachi et al., 2020) and an AUC of up to 0.944 (Deng et al., 2022). The latter approaches ensured subject independence in the training-test procedure and provided an unbiased estimation of results. The use of inertial sensors presented results comparable to those based on force sensors, achieving an accuracy ranging from 0.78 (Stamate et al., 2018) to 0.997 (Lin et al., 2022). On the other hand, the use of RNN, including LSTM (Balaji et al., 2021; Pham, 2021; Xia et al., 2020; Zhao et al., 2018) and Bi-LSTMS (Liu et al., 2021), presented a better performance for gait assessment than the CNN and ANN approaches in most of the studies listed in Table 3.

4.10.3. Tremor

A total of sixteen studies used wearables and DL methods for the assessment of tremor. Eleven studies (Han Byul et al., 2018; Hssayeni, Jimenez-Shahed, Burack, & Ghoraani, 2019; Ibrahim, Zhou, Jenkins, Trejos, & Naish, 2021; Papadopoulos, Kyritsis et al., 2020; Phokaewvarangkul, Vateekul, Wichakam, Anan, & Bhidayasiri, 2021; Qin, Jiang, Chen, Hu, & Ma, 2019; San-Segundo et al., 2020; Sigcha et al., 2021; Tong, He, & Peng, 2021; Varghese et al., 2021; Varghese, Fujarski, Hahn, Dugas, & Warnecke, 2020) focused on the assessment of tremor, while five studies analyzed tremor in combination with symptoms such as dyskinesia (Roy et al., 2013), fine motor impairment (FMI) (Papadopoulos, Iakovakis et al., 2020), cognitive and speech impairment (Lauraitis, Maskeliunas, Damasevicius, & Krilavicius, 2020a), gait (Stamate et al., 2018), and bradykinesia (Lonini et al., 2018; Stamate et al., 2018). The sample information, sensor technology, methods, and results of these studies are summarized in Table 4 and discussed in detail below.

Data. The number of PwPD enrolled in these studies ranged from 1 (Lauraitis et al., 2020a) to 260 (Varghese et al., 2021). All studies listed in Table 4 used proprietary databases in their studies. In twelve studies (Han Byul et al., 2018; Hssayeni et al., 2019; Ibrahim et al., 2021; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; Phokaewvarangkul et al., 2021; Qin et al., 2019; Roy et al., 2013; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018) the MDS-UPDRS scale was used as a reference to assess the severity of the tremor. More specifically, in Ibrahim et al. (2021) 18 PwPD performed MDS-UPDRS tasks to assess resting, postural, and action tremors using IMUs placed on the wrist and hand. In Papadopoulos, Kyritsis et al. (2020), accelerometer measurements of 45 PwPD were recorded in the wild using a smartphone. In Tong et al. (2021), acceleration data was collected from 10 patients while performing scheduled activities (e.g., walking, standing, or turning) using a watch-like IMU. In Qin et al. (2019), surface electromyography (sEMG) was used to acquire data from 147 PwPD. In Sigcha et al. (2021), several MDS-UPDRS tasks, including resting tremor, were recorded from 18 PwPD using a consumer smartwatch accelerometer placed on the dominant hand with tremor. In Han Byul et al. (2018), tremor signals were recorded from 92 PwPD using a custom-developed wrist and finger device. The MDS-UPDRS scale was used to assess the severity of the tremor. In Hssayeni et al. (2019) 24 PwPD performed a series of ADLs wearing a motion sensor placed on the wrist and ankle of the most affected side. In San-Segundo et al. (2020), data were collected in the laboratory and free-living settings of 12 PwPD while performing various ADLs (for example, writing, typing, and playing chess/cards). Data were collected from two smart bands. In Phokaewvarangkul et al. (2021), a glove connected to a mobile application was used to assess, stimulate, and quantify tremors in 20 PwPD. The glove was equipped with electrical muscle stimulation (EMS) to provide muscle stimulation on-demand to suppress tremors. In Lauraitis et al. (2020a) a smartphone with a specific application was used to collect data from 15 subjects (1 PD, 6 with cognitive and motor impairments, 8 HC) while performing cognitive and language tasks. In Stamate et al. (2018), the cloudUPDRS application and smartphones were used to collect data from 12 PwPD while performing guided MDS-UPDRS tasks. In Lonini et al. (2018), 20 PwPD wore adhesive biosensors attached to their skin while performing common tasks (e.g., walking, moving hands, typing). A total of 6 sensors were placed on both hands, thighs, and forearms. In Roy et al. (2013), 19 PwPD wore four sensors attached to each extremity while performing unrestrained activities (e.g., sitting, standing, walking). Data were recorded for 4 h to capture a complete on/off cycle. In Papadopoulos, Iakovakis et al. (2020), data from 40 PwPD were captured by an Android application during user interaction with the device. In Varghese et al. (2021), 260 PwPD performed a predefined neurological examination (duration 15 min) while wearing two smartwatches. In Varghese et al. (2020), acceleration recordings of 192 PwPD were collected from two consumer smartwatches while performing ten different coordination tasks while sitting in a chair.

Pre-processing. Low-pass (Papadopoulos, Kyritsis et al., 2020), high-pass (Han Byul et al., 2018; San-Segundo et al., 2020; Stamate et al., 2018), and band-pass (Hssayeni et al., 2019; Ibrahim et al., 2021; Lonini et al., 2018; Sigcha et al., 2021; Tong et al., 2021) filtering were used to remove low-frequency trends (including motion band) and high-frequency noise in inertial data. In addition, eleven studies used fixed-length windows to segment inertial data into temporal frames, with time windows between 1 s (Tong et al., 2021) and 50 s (Han Byul et al., 2018). Specifically, most studies used window sizes between 2 s and 5 s (Hssayeni et al., 2019; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; Qin et al., 2019; Roy et al., 2013; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018; Tong et al., 2021), while in Han Byul et al. (2018) a longer window (50 s) was used. An overlap of 50% was used in 6 studies (Lonini et al., 2018; Roy et al., 2013; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018; Tong et al., 2021), while two studies did not use any overlap (Hssayeni et al., 2019; Papadopoulos, Iakovakis et al., 2020).

Deep learning models. Ten studies (Han Byul et al., 2018; Ibrahim et al., 2021; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; Qin et al., 2019; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018; Tong et al., 2021) used CNN with raw data input (Ibrahim et al., 2021; Papadopoulos, Kyritsis et al., 2020; Sigcha et al., 2021; Stamate et al., 2018), FFT (Han Byul et al., 2018; Sigcha et al., 2021), and specific features (time and frequency domain) (Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Qin et al., 2019; San-Segundo et al., 2020; Tong et al., 2021). Three studies (Hssayeni et al., 2019; Lauraitis et al., 2020a; Phokaewvarangkul et al., 2021) used LSTM fed with raw data (Phokaewvarangkul et al., 2021) and features (Hssayeni et al., 2019; Lauraitis et al., 2020a). One work used a dynamic neural network (Roy et al., 2013). Finally, two studies used MLP architectures fed with FFT (Varghese et al., 2021) and features in combination with FFT (Varghese et al., 2020). Furthermore, only in a single study (Hssayeni et al., 2019) the best results were obtained using the gradient tree boost algorithm (shallow ML).

Validation. Several validation methods were used to evaluate the performance of the tremor algorithms, including LOSO CV (Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018; Tong et al., 2021), k-fold CV with k equal to 10 (Han Byul et al., 2018; Lauraitis et al., 2020a) or 3 (Phokaewvarangkul et al., 2021), hold-out (Hssayeni et al., 2019; Ibrahim et al., 2021; Qin et al., 2019; Roy et al., 2013) and nested CV with k equal to 5 (Varghese et al., 2021, 2020). The nested CV procedure involves treating the

Table 4
List of studies that used wearable sensors and DL methods for tremor assessment. ACC: accuracy; AUC: area under the curve; Sens: sensitivity; Spec: specificity; MAE: mean absolute error; acc: accelerometer; gyro: gyroscope; mag: magnetometer; CNN: convolutional neural network; LSTM: long short-term memory; ANN: Artificial neural network LOSO: leave-one-subject-out; FFT: fast Fourier transform; CV: cross validation; S-I: subject independent; S-D: subject dependent; RS: random shuffle.

Study	# PD subjects (# Controls)	Sensor type (location)	Model (Input)	Validation type	Performance
Ibrahim et al. (2021)	18	gyro (wrist, finger)	CNN (raw data)	Hold-out (S-I)	ACC 0.992
Papadopoulos, Kyritsis et al. (2020)	31 (14)	acc (hand)	CNN (raw data)	LOSO (S-I)	Sens 0.612 Spec 0.850
Tong et al. (2021)	5 (5)	acc, gyro, mag (wrist)	CNN (features)	LOSO (S-I)	ACC 0.973 Sens 0.915 Spec 0.972
Qin et al. (2019)	147	sEMG (arm)	S-Net (CNN) (features)	Hold-out (S-D)	ACC 0.906 Sens 0.906
Sigcha et al. (2021)	18	acc (wrist)	Multi-task CNN (raw data, FFT)	LOSO (S-I)	AUC 0.936 Sens 0.861 Spec 0.861 r 0.969
Han Byul et al. (2018)	92	acc, gyro (wrist)	CNN (FFT)	K-fold (RS)	ACC 0.85 *r* 0.93
Hssayeni et al. (2019)	24	acc, gyro (wrist, leg)	LSTM (features)	LOSO (S-I)	r 0.77 MAE 1.32
San-Segundo et al. (2020)	12	acc (wrist)	CNN (features)	LOSO (S-I)	AUC 0.884 Spearman 0.9
Phokaewvarangkul et al. (2021)	20 (20)	gyro (wrist)	LSTM (raw data)	K-fold (S-I)	ACC 0.865 Spearman 0.77
Lauraitis et al. (2020a)	1 (8)	touchscreen (hand)	LSTM (Ensemble) (features)	K-fold	ACC 0.961 Sens 0.961 Spec 0.953
Stamate et al. (2018)	12	acc pocket, leg, hand, finger	Recurrent CNN (raw data)	LOSO (S-I)	ACC 0.78 AUC 0.87
Lonini et al. (2018)	20	acc, gyro (hand, arm, leg)	CNN (features)	LOSO (S-I)	AUC 0.79
Roy et al. (2013)	19 (4)	acc, sEMG (arm, leg)	Dynamic Neural Network (features)	Hold-out (S-I)	Sens 0.917 Spec 0.929
Papadopoulos, Iakovakis et al. (2020)	40 (139)	acc, touchscreen (hand)	CNN Ensemble (features)	LOSO (S-I)	AUC 0.87 Sens 0.83 Spec 0.91
Varghese et al. (2021)	260 (89)	acc (wrist)	ANN (FFT, questionary data)	nested CV	ACC 0.823 Sens 0.905
Varghese et al. (2020)	192 (51)	acc (wrist)	ANN (FFT, features)	nested CV	ACC 0.89 Sens 0.92

optimization of the model's hyperparameters as part of the model itself and evaluating it within the k-fold. In practice, nested CV involves the implementation of a double loop, an outer loop that is used to evaluate the quality of the model, and an inner loop that is used for parameter selection.

Results. In one study (Hssayeni et al., 2019), a regression model was used to predict the clinical reference scale (UPDRS). In this work, Pearson's correlation coefficient r of 77 and an MAE of 1.32 were reported to evaluate the performance of the model. The other fifteen studies performed classification tasks to detect the presence and severity of tremor, with two studies (Han Byul et al., 2018; Sigcha et al., 2021) reporting both classification metrics (accuracy 0.85 Sigcha et al., 2021, AUC 0.936 Han Byul et al., 2018) and strong correlations ($r >$ 0.93) with the clinical score. Most studies reported results in terms of accuracy, with values ranging from 0.78 (Stamate et al., 2018) to 0.992 (Ibrahim et al., 2021). Furthermore, the sensitivity ranged from 0.612 (Papadopoulos, Kyritsis et al., 2020) to 0.946 (Lauraitis et al., 2020a) and the specificity from 0.850 (Papadopoulos, Kyritsis et al., 2020) to 0.972 (Tong et al., 2021). Furthermore, five studies (Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; San-Segundo et al., 2020; Sigcha et al., 2021; Stamate et al., 2018) provided an AUC ranging from 0.79 (Lonini et al., 2018) to 0.936 (Sigcha et al., 2021). Regarding the DL models, on average, CNN performed better in terms of accuracy than the LSTM and MLP approach, while LSTM can provide a high correlation (r 0.77) (Hssayeni et al., 2019) compared to UPDRS.

4.10.4. Bradykinesia

A total of twelve studies used wearable sensors and DL methods for the detection and/or monitoring of bradykinesia, including the evaluation of FMI (Iakovakis et al., 2020; Papadopoulos, Iakovakis et al., 2020; Prince et al., 2019). Of these studies, two focused on the evaluation of bradykinesia (Borzì et al., 2020; Park et al., 2021), while ten studies analyzed bradykinesia in combination with symptoms such as dyskinesia (Pfister et al., 2020), tremor (Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Stamate et al., 2018), gait (Deng et al., 2022; Prince et al., 2019; Stamate et al., 2018), dysphonia (Deng et al., 2022; Oung, Muthusamy, Basah, Lee, & Vijean, 2018; Prince

et al., 2019; Shichkina, Stanevich, & Irishina, 2020), and rigidity (Iakovakis et al., 2020). Information on the sample, sensor technology, methods, and results is summarized in Table 5 and discussed below.

Data. The sample size ranged from 10 (Shichkina et al., 2020) to 698 (Deng et al., 2022) PwPD. In nine studies, proprietary databases were used, while in Hssayeni, Jimenez-Shahed, Burack et al. (2021) a proprietary database (ADL-like tasks) was used in addition to the Physical Activity Monitoring Data Set (PAMAP2) for transfer learning tasks. In Deng et al. (2022) and Prince et al. (2019), data were collected using the mPower application. Patients were asked to touch the touchscreen of the smartphone as fast as possible. The acceleration and position of the touch on the screen were recorded during the experiments. In more detail, in Prince et al. (2019) data were collected from 799 PwPD and 534 HC while performing the finger touch task (Bot et al., 2016). In Deng et al. (2022) data from the DREAM Parkinson's disease challenge (mPower) (Sieberts et al., 2021) were used. Data used for the evaluation of FMI included data (finger tapping activity) from 1057 PwPD and 5343 HC. In Shichkina et al. (2020), data were collected through a smartphone application when subjects performed a wide range of tests, also recording background data (sensor data). In Park et al. (2021), 25 PwPD and 21 HC were evaluated while performing three selected elements of Part III of MDS-UPDRS (e.g., finger tapping, hand movements, and alternating rapid movements), with two gyroscopes on the thumb and index finger recording data. In Borzì et al. (2020), 93 PwPD were fitted with a smartphone on their thigh and had to perform the MDS-UPDRS leg agility task (Item 3.8). In Stamate et al. (2018), a smartphone application was developed to guide patients in performing certain tasks and to collect and analyze data. In particular, hand pronation–supination, leg agility, and finger touch were analyzed to detect bradykinesia. In Lonini et al. (2018), 20 PwPD were equipped with 6 flexible wearable sensors (with accelerometer and gyroscope) attached to their hands, arms, and thighs. They were instructed to perform various tasks, including some ADL (e.g., walking, sitting upright), items related to MDS-UPDRS (e.g., finger on nose, alternating hand movements), and other more specific tasks (e.g., organizing a series of folders, folding towels, pouring water from a bottle, and drinking). In Oung et al. (2018), accelerometer and gyroscope recordings were recorded from four IMUs placed on the hands and upper limbs of 65 participants. Subjects were asked to perform a series of standardized activities, including getting up from a chair, moving their hands, tapping their fingers, touching their feet, and moving their legs. Finally, in Iakovakis et al. (2020) data from typing sessions were collected using a touchscreen using a mobile application (iPrognosis).

Pre-processing. Low-pass (Borzì et al., 2020; Stamate et al., 2018) and band-pass filters (Hssayeni, Jimenez-Shahed, Burack et al., 2021; Lonini et al., 2018; Pfister et al., 2020; Stamate et al., 2018) were used to remove low-frequency trends and high-frequency noise, while in Park et al. (2021) signal integration was used to convert angular velocity data into an angle to track the movement of the patient's hand. Two studies (Borzì et al., 2020; Park et al., 2021) analyzed the data collected from the entire task, while the other studies divided the original signal into time windows of equal duration. Specifically, sliding windows of 5 s (Hssayeni, Jimenez-Shahed, Burack et al., 2021; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Stamate et al., 2018), 10 s (Oung et al., 2018; Park et al., 2021), and 60 s (Pfister et al., 2020) were used in the inertial signal segmentation procedure.

Deep learning models. Seven studies used CNNs (Deng et al., 2022; Iakovakis et al., 2020; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Pfister et al., 2020; Prince et al., 2019; Stamate et al., 2018), including a recurrent CNN (Stamate et al., 2018), a CNN ensemble (Papadopoulos, Iakovakis et al., 2020), and an autoencoder CNN (Iakovakis et al., 2020). Among these models, four networks were fed with raw signals (Deng et al., 2022; Lonini et al., 2018; Pfister et al., 2020; Stamate et al., 2018), while three were fed with features (Iakovakis et al., 2020; Papadopoulos, Iakovakis et al., 2020; Prince et al., 2019). In one study (Oung et al., 2018), EWT was calculated and used as input for a probabilistic neural network (PNN). In other studies, a set of features extracted from the temporal and spectral domains was fed by an ANN (Borzì et al., 2020; Park et al., 2021), a deep ensemble (Hssayeni, Jimenez-Shahed, Burack et al., 2021), and a GRU model (Shichkina et al., 2020).

Validation procedure. Different validation methods were used in these studies. Four articles used k-fold CV, with k equal to 5 (Deng et al., 2022; Park et al., 2021) and 10 (Oung et al., 2018; Prince et al., 2019). The other approaches included hold-out (Iakovakis et al., 2020; Shichkina et al., 2020) and LOSO CV (Borzì et al., 2020; Hssayeni, Jimenez-Shahed, Burack et al., 2021; Lonini et al., 2018; Papadopoulos, Iakovakis et al., 2020; Pfister et al., 2020; Stamate et al., 2018). All these validation methods guarantee the independence of the subjects in the training-test procedure and provide an unbiased estimation of the results.

Results. In two studies (Borzì et al., 2020; Hssayeni, Jimenez-Shahed and Burack, 2021), regression models were used to predict the clinical reference scale for the specific symptom, while the remaining works aimed to classify the presence or absence of the symptom. Pearson's correlation coefficient and RMSE (Borzì et al., 2020) were used to evaluate the performance of the regression model, while AUC, accuracy, sensitivity, and specificity were used to evaluate the performance of classification algorithms. In Borzì et al. (2020), a strong correlation ($r = 0.92$) was obtained by comparing the prediction of the model with the mean severity of bradykinesia assessed by four experienced clinicians. Accuracy of up to 0.89 (Shichkina et al., 2020) was reported and the AUC ranged from 0.77 (Lonini et al., 2018) to 0.93 (Park et al., 2021) for the detection of bradykinesia. When bradykinesia was assessed using data collected from finger or toe tapping activities, the combination of features and ANN showed very good results, with an AUC greater than 0.92 (Borzì et al., 2020; Park et al., 2021), slightly lower than CNN fed raw data (Deng et al., 2022). Evaluation of additional activities (e.g., walking (Lonini et al., 2018), postural transitions (Oung et al., 2018)) did not provide an improvement in performance.

4.10.5. Dyskinesia

Three studies used the combination of wearables and DL models for the detection and/or monitoring of dyskinesia. Of this group, one study focused on dyskinesia (Hssayeni, Jimenez-Shahed, Burack, 2021), while the other two studies also analyzed bradykinesia (Pfister et al., 2020) and tremor (Roy et al., 2013) in addition to dyskinesia. Information on the sample, sensor technology, methods, and results is summarized in Table 6 and discussed below.

Data. In all studies, proprietary data sets were used, with sample sizes ranging from 15 (Hssayeni, Jimenez-Shahed, Burack, 2021) to 30 (Pfister et al., 2020) PwPD. In Hssayeni, Jimenez-Shahed, Burack (2021), 15 PwPD wore two inertial sensors on the wrist and ankle of the most affected side and performed four sets of ADLs under different treatment conditions (e.g., OFF, ON, and after a certain period of time after taking the medication). Activities included walking, resting, and various other ADLs (e.g., using a knife and fork to cut food, putting on and taking off a coat, and drinking water from a cup). In Pfister et al. (2020), inertial data from 30 PwPD were recorded using an IMU sensor worn on the wrist of the most affected side. Various activities were recorded in free-living conditions, including sitting, lying down, walking, standing, and sleeping. In Roy et al. (2013), 19 PwPD were equipped with a wearable sensor (incorporating an accelerometer and sEMG) on each limb (e.g., two on the forearms and two on the lower legs). Data were recorded in a simulated home environment while subjects performed simple (e.g., sitting, standing, walking, lying down) and complex (e.g., preparing snacks, eating, reading, writing) ADLs.

Pre-processing. Band-pass filtering (Hssayeni, Jimenez-Shahed, Burack, 2021; Pfister et al., 2020) was used to remove low-frequency trends and high-frequency noise. Three studies segmented the recorded signals into data windows of 2 s (Roy et al., 2013), 5 s (Hssayeni,

Table 5
List of studies that used wearable sensors and DL methods for the detection or monitoring of bradykinesia. ACC: accuracy; Sens: sensitivity; Spec: specificity; AUC: area under the curve; ; RMSE: root mean square error; MAE: mean absolute error; FMI: fine motor impairment; acc: accelerometer; gyro: gyroscope; mag: magnetometer; CNN: convolutional neural network; GRU: gated recurrent unit; ANN: artificial neural network; RCNN: recurrent CNN; PNN: probabilistic neural network; EWT: empirical wavelet transform; LOSO: leave-one-subject-out; S-I: subject independent; S-D: subject dependent; RS: random shuffle.

Study	# PD subjects (# Controls)	Sensor type (position)	Model (Input)	Validation	Performance
Prince et al. (2019)	799 (524)	acc, touchscreen (waist, hand)	CNN (features)	k-fold (stratified)	ACC 0.707 (FMI) ACC 0.774 (gait)
Borzì et al. (2020)	93	acc, gyro (thigh)	ANN (features)	LOSO (S-I)	r 0.92 RMSE 0.42 AUC 0.920
Hssayeni, Jimenez-Shahed, Burack et al. (2021)	24 (9)	gyro (wrist, leg)	Deep Ensemble (features)	LOSO (S-I)	r 0.79 MAE 5.95
Shichkina et al. (2020)	10 (18)	acc, touchscreen	GRU (features)	Hold-out (S-D)	ACC 0.890
Pfister et al. (2020)	30	acc (wrist)	CNN (raw data)	LOSO (S-I)	Sens 0.645 Spec 0.892 ACC 0.654
Park et al. (2021)	25 (21)	gyro (index, thumb)	ANN (features)	k-fold	AUC 0.926
Stamate et al. (2018)	12	acc (variable)	RCNN (raw data)	LOSO (S-I)	AUC 0.870 F-score 0.820
Lonini et al. (2018)	20	acc, gyro (hands, forearm, thighs)	CNN (raw data)	Hold-out (S-I)	AUC 0.776
Papadopoulos, Iakovakis et al. (2020)	40 (139)	acc, touchscreen (hand)	CNN Ensemble (features)	LOSO (S-I)	Sens 0.83 Spec 0.91 AUC 0.87
Oung et al. (2018)	50 (15)	acc, gyro, mag (wrists, legs)	PNN (EWT features)	k-fold (RS)	ACC 0.899
Deng et al. (2022)	698 (4329)	acc, touchscreen (hand)	CNN (raw data)	k-fold (S-I)	AUC 0.944
Iakovakis et al. (2020)	22 (17)	touchscreen (hand)	CNN Autoencoder (features)	Hold-out (S-I)	AUC 0.97 (UPDRS 23) AUC 0.93 (UPDRS 31)

Table 6
List of studies that used wearable sensors and DL methods for the detection or monitoring of dyskinesia. ACC: accuracy; Sens: sensitivity; Spec: specificity; MAE: mean absolute error; acc: accelerometer; gyro: gyroscope; sEMG: surface electromyography; CNN: convolutional neural network; BiLSTM: Bidirectional long short-term memory network; ANN: artificial neural network; S-I: subject independent; LOSO: leave-one-subject-out; LOTO: leave-one-task-out.

Study	# PD subjects (# Controls)	Sensor type (position)	Model (Input)	Validation	Performance
Pfister et al. (2020)	30	acc (wrist)	CNN (raw data)	LOSO (S-I)	ACC 0.654 Sens 0.64 Spec 0.89
Roy et al. (2013)	19 (4)	acc, sEMG (forearm, leg)	ANN (features)	Hold-out (S-I)	Sens 0.917 Spec 0.895
Hssayeni, Jimenez-Shahed, Burack (2021)	15	acc, gyro (wrist, ankle)	BiLSTM (features)	LOTO (S-I)	r 0.87 MAE 1.74 Sens 0.850 Spec 0.850

Table 7
List of studies using wearable sensors and DL methods to detect other motor symptoms. acc: accelerometer; gyro: gyroscope; mag: magnetometer; CNN: convolutional neural network; ANN: artificial neural; ACC: accuracy; Sens: sensitivity; AUC: area under the curve; S-I: subject independent.

Study	# PD subjects (# Controls)	Sensor type (position)	Model (Input)	Validation	Performance
Moon et al. (2020)	524	acc, gyro, mag (wrist, feet, chest, waist)	ANN (features)	k-fold (stratified)	ACC 0.89 Sens 0.61 F-score 0.61
Iakovakis et al. (2020)	22 (17)	touchscreen (hand)	CNN Autoencoder (features)	Hold-out (S-I)	AUC 0.97 (MDS-UPDRS 22)

Jimenez-Shahed, Burack, 2021), and 60 s (Pfister et al., 2020), with (Pfister et al., 2020; Roy et al., 2013) or without (Hssayeni, Jimenez-Shahed, Burack, 2021) overlap.

Deep learning models. Two out of three studies used extracted features to feed a BiLSTM network (Hssayeni, Jimenez-Shahed, Burack, 2021) and an ANN (Roy et al., 2013). A single study employed a CNN model fed with raw data.

Validation procedure. Several validation methods were used, including 5-fold CV (Pfister et al., 2020), hold-out (Roy et al., 2013), and subject-dependent leave-one-task-out (LOTO) CV (Hssayeni, Jimenez-Shahed, Burack, 2021). Unlike the first two approaches, the latter takes advantage of the availability of several recording sessions of each participant to perform subject-dependent classification. Specifically, it consists of training the classification/regression model with data from all but one session, which is used as a test.

Results. In one study (Hssayeni, Jimenez-Shahed, Burack, 2021), a regression model was used to predict dyskinesia severity using the clinical reference scale. In this case, Pearson's correlation coefficient and MAE were used to evaluate the model's performance. The other two studies (Pfister et al., 2020; Roy et al., 2013) performed classification tasks to detect the presence or absence of dyskinesia. Accuracy (Pfister et al., 2020), sensitivity, and specificity (Roy et al., 2013) were used to assess the performance of the classification models. The LSTM prediction was strongly correlated ($r = 0.87$) with the modified involuntary abnormal movements scale (mAIMS) in Hssayeni, Jimenez-Shahed, Burack (2021). Sensitivity ranged from 0.64 (Pfister et al., 2020) to 0.92 (Roy et al., 2013) and specificity from 0.89 (Hssayeni, Jimenez-Shahed, Burack, 2021) to 0.90 (Pfister et al., 2020; Roy et al., 2013). ANN fed with extracted features (Roy et al., 2013) showed superior results compared to other DL methods (Hssayeni, Jimenez-Shahed, Burack, 2021; Pfister et al., 2020).

4.10.6. Other motor symptoms

Two studies also analyzed other motor aspects, including balance (Moon et al., 2020) and rigidity (Iakovakis et al., 2020). The first (Moon et al., 2020) analyzed gait impairment in addition to balance, while the second (Iakovakis et al., 2020) analyzed bradykinesia and FMI in addition to rigidity. Sample information, sensor technology, methods, and results are summarized in Table 7 and discussed below.

Data. In studies investigating balance (Moon et al., 2020) and rigidity (Iakovakis et al., 2020), proprietary databases were used with 524 and 22 PwPD, respectively. In more detail, in Moon et al. (2020), PwPD (n = 524) and those with essential tremor (n = 43) were equipped with six IMUs and asked to perform standing and walking tests. From these data, several balance and gait variables were extracted from inertial signals. In Iakovakis et al. (2020), data from typing sessions, including temporal information associated with key presses and key releases (key dynamics), were collected using a touchscreen and a mobile application (iPrognosis). These data were compared with the stiffness score using the MDS-UPDRS Part III item 22 guidelines.

Pre-processing. Various preprocessing techniques were applied to analyze movement characteristics (Moon et al., 2020) and sequences of key-stroke dynamics (Iakovakis et al., 2020). In Moon et al. (2020), the set of balance and gait features was balanced using the SMOTE technique. In Iakovakis et al. (2020), a conditional filter was applied to discard outliers (e.g., prolonged presses) from the key dynamics sequences. In addition, zero padding was applied to fill in smaller sequences.

Deep learning models. In Moon et al. (2020), an ANN architecture was used, fed with balance and gait features extracted using commercial software (Moon et al., 2020). In Iakovakis et al. (2020), CNN coders were used to analyze the temporal information associated with key dynamics.

Validation. The k-fold stratification was used in Moon et al. (2020) with a k equal to 3. This CV procedure was used to maintain the sample rate for each class during training and evaluation. In Iakovakis et al. (2020), holdout-out validation was employed in a subject-independent manner, ensuring the independence of subject data in training-test subsets.

Results. In Moon et al. (2020), classification was used to distinguish between PD and subjects with essential tremor. In this study, the use of ANN and movement features provided the best results in terms of accuracy (0.89), sensitivity (0.61), and F-score (0.61) compared to classical ML algorithms. In Iakovakis et al. (2020), the best results in the diagnosis of PD (AUC 0.97) were obtained using the predicted stiffness indices obtained from the CNN autoencoder fed with key dynamics data (see Table 7).

4.11. Non-motor symptoms

Fifteen studies employed DL approaches to assess NMM. This section describes technological solutions based on wearable sensors and DL methods to assess different NMM of PD, including speech impairment (Section 4.11.1) and cognitive impairment (Section 4.11.2). A summary of these symptoms is presented in the following subsections, where the data set used, signal processing techniques, DL methods, validation methodologies, and results are reported and briefly discussed.

4.11.1. Speech impairment

The review of the literature conducted in this study revealed eleven articles aimed at evaluating language impairment in PwPD through the analysis of vocal samples recorded using smartphones and professional microphones. Table 8 reports the information derived from the identified works, including sample size, sensor technology, methods, and results.

Data. Eleven articles (Berus, Klancnik, Brezocnik, & Ficko, 2018; Deng et al., 2022; Prince et al., 2019; Wan, Liang, Zhang, & Guizani, 2018; Xiong & Lu, 2020; Zhang, 2017; Zhang, Wang, Zhang, Jin, & Zhang, 2021) used freely available databases, including mPower (Bot et al., 2016) and feature sets from University of California Irvine (UCI) machine learning archives (Little et al., 2007; Sakar et al., 2013, 2019; Tsanas et al., 2009). In more detail, the authors in Prince et al. (2019) used a subset of data from 1072 subjects (679 PwPD and 393 HC) from the mPower database (Sage Bionetworks, 2016), a crowd-sourced corpus in which enrolled volunteers performed the required tests through an iPhone application. All samples were collected unsupervised and the expert's diagnosis was reported by the participants themselves. In this corpus, the speakers had to perform various motor tasks and record the vowel */a/* in a sustained manner for 10 s. Similarly, in Deng et al. (2022) recordings of the DREAM challenge on Parkinson's disease (mPower) were used to compare FMI diagnosis with voice-based diagnosis models. The data used for the voice analysis corresponded to recordings of the vowel */a/* (sustained mode) of 645 PwPD and 2084 HC. In Berus et al. (2018), Wan et al. (2018), Xiong and Lu (2020) and Zhang (2017), the UCI datasets provide pre-extracted feature vectors calculated from samples recorded in different languages, recording conditions, and techniques. In more detail, the authors in Xiong and Lu (2020) performed their analysis on the *Parkinson's Disease Classification Data Set* collected in Sakar et al. (2019), which includes recordings of 188 PwPD and 64 HC pronouncing the vowel */a/* in a sustained manner. In Wan et al. (2018), the authors' main objective was to estimate the severity of PD using smartphones, so they conducted their analysis on the *Parkinson's Telemonitoring Data Set* collected in Tsanas et al. (2009), which includes features extracted from sustained vowel phonations of 42 PwPD at an early stage of the disease engaged in a six-month telemonitoring tool. The authors of Zhang (2017) used two different UCI datasets for their analysis: the first *Parkinson's Data Set* collected in Little et al. (2007) and consisting of features extracted from voice measurements of 31 individuals (23 PD), and the second *Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set* collected in Sakar et al. (2013) which includes 20 PwPD and 20 HC. This second corpus was also used in Berus et al. (2018), whose

authors analyzed not only the features derived from sustained vowels, but also nine Turkish words and a list of numbers. Finally, four studies (Lauraitis, Maskeliūnas, Damaševičius, & Krilavičius, 2020b; Oung et al., 2018; Shichkina et al., 2020; Zhang et al., 2021) used proprietary databases for their analyzes. In more detail, the authors of Lauraitis et al. (2020b) collected 339 voice samples from 7 patients with neurological disorders (1 with PD) and 8 HC in 5 face-to-face visits. The task is part of a mobile application proposed by the authors that guides users in reading a short text of a predefined poem and records using the microphone of the mobile device. Similarly, the data set used for the analysis in Shichkina et al. (2020) was derived from voice samples collected through a mobile application. As reported in the study, 28 subjects (10 with PD) were enrolled in the data collection procedure and asked to pronounce the vowel */a/* for as long as possible. In Oung et al. (2018), voice recordings were recorded using a headset microphone when subjects sustained the vowel */a/* for as long as possible, trying to keep frequency and amplitude stable. Finally, (Zhang et al., 2021) explored the learning of multivariate time series representations for the diagnosis of chronic diseases. For this purpose, the authors used two different data sets, the first consisting of 48 PwPD and 20 HC pronouncing different types of sound recordings, and the second comprising a heterogeneous data source consisting of advanced imaging, biological sampling, and clinical and behavioral assessments of 466 PwPD and 217 HC.

Pre-processing. Most of the reviewed studies did not report comprehensive information on the pre-processing techniques applied to the input signals, also due to the presence of pre-extracted feature sets. Regarding window lengths and overlap, only the authors of Lauraitis et al. (2020b) reported the values adopted for the segmentation of speech signals into time intervals (e.g., window length = 0.0052 ·Fs, overlap = 0.0042 ·Fs). Five studies (Berus et al., 2018; Lauraitis et al., 2020b; Prince et al., 2019; Shichkina et al., 2020; Wan et al., 2018) applied a normalization step, which in Wan et al. (2018) was also combined with a low-pass filter to de-noise the signal. In addition, the authors of Lauraitis et al. (2020b) and Oung et al. (2018) applied silence removal using a threshold approach (Lauraitis et al., 2020b) or endpoint detection (Oung et al., 2018), while (Zhang et al., 2021) used a filler algorithm to have predefined signal lengths before feature extraction.

Deep learning models. Two studies employed RNN, including BiLSTM (Lauraitis et al., 2020a, 2020b) and GRU (Shichkina et al., 2020). On the other hand, two studies (Berus et al., 2018; Wan et al., 2018) used ANN, and two studies (Deng et al., 2022; Prince et al., 2019) proposed CNN models. Regarding (Xiong & Lu, 2020; Zhang, 2017; Zhang et al., 2021), the use of autoencoders for automatic feature extraction was proposed. However, in Xiong and Lu (2020) and Zhang (2017) the features extracted by the autoencoder were used in conjunction with the linear discriminant analysis (LDA) and k-nearest neighbors (kNN) classifiers, while in Zhang et al. (2021) the use of MLP layers was proposed for classification tasks. Finally, in a single study (Oung et al., 2018), the EWT was calculated and used as input for a PNN.

Validation. Two studies (Berus et al., 2018; Wan et al., 2018) used LOSO validation, which guarantees speaker independence in the case of multiple recordings of the same individual. The k-fold CV approach was employed in Deng et al. (2022), Lauraitis et al. (2020a), Oung et al. (2018), Prince et al. (2019) and Xiong and Lu (2020). In Lauraitis et al. (2020a), Oung et al. (2018), Prince et al. (2019) and Xiong and Lu (2020), the number of folds was set to 10, while in Deng et al. (2022) the number of folds was set to 5. Furthermore, as the authors of Lauraitis et al. (2020a), Oung et al. (2018), Prince et al. (2019) and Xiong and Lu (2020), the number of folds was fixed at 10, while in Deng et al. (2022), the number of folds was fixed at 5. Furthermore, as the authors of Prince et al. (2019) demonstrated the possibility of predicting PD in the presence of missing source data. Specifically, the data set was previously divided into two subgroups comprising individuals with missing and complete source data, to be used during the training/validation and testing phases, respectively. Finally, four of the reviewed articles used a hold-out approach with different percentages of subjects in training and testing sets: 70%–30% in Lauraitis et al. (2020b), 80%–20% in Shichkina et al. (2020), 50%–50% in Zhang (2017), and 90%–10% in Zhang et al. (2021).

Results. Accuracy was used mainly to measure the performance of models to diagnose, predict, and monitor PD through speech analysis. In more detail, accuracy values ranging from 0.8 (Wan et al., 2018) to 0.98 (Zhang, 2017) were reported. In Prince et al. (2019), the authors preferred the F-score to the accuracy metric due to the unbalanced distribution of the data set across classes, reporting an F-score of 0.850. Furthermore, the authors used sensitivity and specificity widely to describe the predictive capacity of the models. Regarding sensitivity, it ranged from 0.73 (Wan et al., 2018) to 0.969 (Zhang et al., 2021), while specificity ranged from 0.725 (Wan et al., 2018) to 0.92 (Xiong & Lu, 2020). A single paper (Deng et al., 2022) reported an AUC of 0.834. It is worth noting that Lauraitis et al. (2020a, 2020b), O'Day et al. (2022), Shichkina et al. (2020) and Zhang (2017) only reported accuracy values.

4.11.2. Cognitive aspects

Five studies used wearable technology and DL to investigate cognitive aspects of PD, both for diagnostic and monitoring purposes. Two studies focused on the assessment of brain function (AlZubi et al., 2020; Oh et al., 2020), two on the assessment of cognitive health (Lauraitis et al., 2020a; Prince et al., 2019), and a single study (Dar et al., 2022) investigated emotional dysfunction. Table 9 shows information from the identified works, including sample size, sensor technology, methods, and results.

Data. Three studies (AlZubi et al., 2020; Lauraitis et al., 2020a; Oh et al., 2020) used proprietary databases, while (Prince et al., 2019) used the mPower (Sage Bionetworks, 2016) database, and Dar et al. (2022) incorporated proprietary data and a combination of publicly available data sets. More specifically, in Prince et al. (2019) memory activity data were used to assess NMM of PD. Data were collected from 39 participants (25 PwPD and 14 HC), who were asked to recall a drawing based on a grid of flowers that appeared on an iPhone screen. In Oh et al. (2020), the authors proposed an EEG-based DL approach to identify PD, based on proprietary data. Data collection involved 40 subjects (20 PwPD and 20 age-matched HC). For this task, 5 min of EEG in the resting state were recorded through a 14-channel EEG headset. In AlZubi et al. (2020), a proprietary database of patients' brain activity was collected to predict changes in brain function. Data were collected from 16 PwPD using a deep brain simulation (DBS) sensor. During this process, the electrodes of the DBS device were connected to a small (wearable) stimulator/sensor placed at the chest. Neural and cognitive deficits were also studied in Lauraitis et al. (2020a), based on self-administered cognitive tests. The self-administered gerocognitive examination (SAGE) was digitized in a mobile application; data were collected from 7 subjects with neurological disorders (1 PwPD, 6 subjects with cognitive and motor deficits) and 8 HC. Finally, an EEG-based approach was also adopted in Dar et al. (2022), to classify emotional states in PD. Two public data sets (AMIGOS, SEED-IV) were incorporated into a proprietary data set for PD. The AMIGOS and PD data sets contained labels for six basic emotion categories (i.e. happiness, sadness, surprise, fear, anger, and disgust), while the SEED-IV included the classes neutral, sad, fear, and happy. Similarly to the PD data set, the AMIGOS database consisted of a 14-channel EEG recorded through a headset. In contrast, the SEED-IV data set comprises 64-channel EEG recordings, from which only 14 channels were selected for standardization.

Pre-processing. Band-pass filtering and threshold-based artifact removal were used for EEG data analysis (Oh et al., 2020). Other processing techniques included normalization in DBS (AlZubi et al., 2020) and EGG (Dar et al., 2022) data. In Lauraitis et al. (2020a) and Prince et al. (2019), no preprocessing was reported for data collected from

Table 8
List of studies using wearable sensors and DL methods for the evaluation of speech impairment. SP: smartphone; ACC: accuracy; Sens: sensitivity; Spec: specificity; EWT: empirical wavelet transform; CNN: convolutional neural network; PNN: probabilistic neural network; GRU: gated recurrent unit; BiLSTM: Bidirectional long short-term memory network; LOSO: leave-one-subject-out; S-I: subject independent; S-D: subject dependent; RS: random shuffle.

Study	# PD subjects (# Controls)	Sensor type	Model (Input)	Validation type	Performance
Prince et al. (2019)	679 (393)	SP microphone	CNN (features)	k-fold (stratified)	ACC 0.782 F-1 score 0.850
Lauraitis et al. (2020b)	1 (8)	SP microphone	BiLSTM (features)	Hold-out (S-D)	ACC 0.945
Xiong and Lu (2020)	188 (64)	microphone	AutoEncoder (features)	k-fold	ACC 0.91 Sens 0.94 Spec 0.92
Wan et al. (2018)	62 (20)	SP microphone	ANN (features)	LOSO (S-I)	ACC 0.80 Sens 0.73 Spec 0.725
Shichkina et al. (2020)	10 (18)	SP microphone	GRU (features)	Hold-out (S-D)	ACC 0.890
Zhang (2017)	71 (28)	microphone	AutoEncoder (features)	Hold-out (S-D)	ACC 0.98
Lauraitis et al. (2020a)	15 (8)	SP microphone	BiLSTM (features)	K-fold (RS)	ACC 0.943
Oung et al. (2018)	50 (15)	microphone	PNN (EWT features)	k-fold (RS)	ACC 0.911
Deng et al. (2022)	645 (2084)	SP microphone	CNN (raw data)	k-fold (S-I)	AUC 0.834
Zhang et al. (2021)	514 (237)	microphone	AutoEncoder (features)	Hold-out (RS)	ACC 0.946 Sens 0.969
Berus et al. (2018)	20 (20)	microphone	ANN (features)	LOSO (S-I)	ACC 0.865 Sens 0.889 Spec 0.840

Table 9
List of studies that used wearable sensors and DL methods for cognitive aspects. SP: smartphone; ACC: accuracy; Sens: sensitivity; Spec: specificity; EEG: Electroencephalography; DBS; Deep brain simulation; ANN: artificial neural; CNN: convolutional neural network; HTSMNN: Heuristic tubu optimized sequence modular neural network; LSTM: Long short-term memory network; S-I: subject independent.

Study	# PD subjects (# Controls)	Sensor type	Model (Input)	Validation type	Performance
Prince et al. (2019)	25 (14)	SP touch screen	ANN Ensemble (features)	k-fold (stratified)	ACC 0.692 F-1 score 0.790
Oh et al. (2020)	20 (20)	EEG headset	CNN (raw data)	k-fold (stratified)	ACC 0.883 Sens 0.847 Spec 0.918
AlZubi, Alarifi, and Al-Maitah (2020)	16	DBS	HTSMNN (features)	Hold-out	ACC 0.982 Sens 0.978
Lauraitis et al. (2020a)	1 (8)	SP touch screen	LSTM (Ensemble) (features)	k-fold	ACC 0.961 AUC 0.983 Sens 0.961 Spec 0.953 F1-score 0.96
Dar, Akram, Yuvaraj, Gul Khawaja, and Murugappan (2022)	20 (68)	EEG headset	CNN+LSTM (features)	LOSO (S-I)	ACC 0.751

smartphone screens. In more detail, in Dar et al. (2022) EEG data from the SEED-IV data set were resampled at 200 Hz and PD data were further resampled at 128 Hz (Dar et al., 2022). The authors of Oh et al. (2020) used a band-pass filter (1–49 Hz) for EEG and applied a threshold technique (100 $mmuV$) to discard artifacts due to blinking. The EEG signals were segmented into 2 s windows for further analysis. In Dar et al. (2022), all EEG recordings were referred to the common mean, and the first 5 s of each EEG recording were considered as baseline activity. Baseline removal was performed in 1 s mini-epochs by subtracting the mean baseline value from the raw EEG activity in each 1 s mini-epoch. The mini-epochs were further standardized using z-score normalization. Frequency bands (i.e. delta, theta, alpha, beta, and gamma) were extracted from each mini-epoch through a filter and used as a feature vector for the DL architecture. Topological scalp maps were also constructed to observe, for each brain region, the contribution of different frequency bands in the elicitation of emotional states. Little information on the preprocessing of the SAGE questionnaire scores was provided in Lauraitis et al. (2020a).

Deep learning models. A single study (Prince et al., 2019) used an ensemble consisting of an ANN, logistic regression (LR), and random forest (RF) fed by features. In Oh et al. (2020), a CNN fed with raw multichannel EEG data was implemented. In AlZubi et al. (2020), the use of a sequence-optimized modular neural network (kHTSMNN) was proposed, fed with features. This network decomposes the input features into small tasks to reduce complexity; furthermore, the number of hidden layers was dynamically determined during the analysis process. In Lauraitis et al. (2020a), an ensemble method was implemented comprising both DNN such as LSTM, and shallow models. Finally, in Dar et al. (2022) the authors employed a CRNN consisting of a CNN, an LSTM, and an MLP.

Validation. In four studies, a k-fold CV was adopted during training, with k set to 10 (Lauraitis et al., 2020a; Oh et al., 2020; Prince et al., 2019) and 20 (Dar et al., 2022). Furthermore, stratified approaches were used in Oh et al. (2020) and Prince et al. (2019) to divide the data. A single study (AlZubi et al., 2020) used the hold-out validation method, with a split ratio of 80/20% for the training and test set, respectively. Furthermore, in Dar et al. (2022) the authors incorporated LOSO CV to assess the generalizability of the proposed architecture.

Results. All studies reported overall accuracy as a metric to measure model performance, with values ranging from 0.692 (Prince et al., 2019) to 0.982 (AlZubi et al., 2020). An AUC of 0.983 was reported in Lauraitis et al. (2020a). In both (Lauraitis et al., 2020a; Oh et al., 2020), sensitivity and specificity metrics were also provided, with values of 0.847 and 0.918 in Oh et al. (2020) and 0.961 and 0.953 in Lauraitis et al. (2020a), respectively. The authors also reported the F-score with values of 0.790 in Prince et al. (2019) and 0.960 in Lauraitis et al. (2020a).

4.12. Results summary

The results of the review indicate that the combination of DL and wearables presents attractive solutions to objectively assess motor symptoms and NMM in PD. The use of wearables allows for the collection of different data modalities (e.g., inertial signals, bioelectrical signals, speech recordings, questionnaire data, etc.) in an economic and non-intrusive fashion. In addition, the application of DL techniques brings powerful methods for processing these data in their raw forms, reducing the efforts of pre-processing and feature engineering tasks (i.e., filtering, data transformation, feature extraction, and feature selection). A summary of application areas, wearable technology, DL algorithms, validation methodologies, data source, and algorithmic approaches reported in the selected studies is presented below.

Application area. In total, sixty-nine studies were identified that used a combination of wearables and DL to assess motor symptoms and NMM in PD. The most common application area was disease monitoring (n = 41, 60%), followed by computer-assisted diagnosis (n = 25, 37%), while only three studies (4%) were identified to focus on predicting response to treatment.

Symptoms. Fifty-nine studies (85%) focused on the analysis of motor symptoms, while ten studies (15%) addressed NMM. Among the studies considered in this review, only five studies (Deng et al., 2022; Lauraitis et al., 2020a; Oung et al., 2018; Prince et al., 2019; Wan et al., 2018) studied both motor symptoms and NMM simultaneously. Regarding motor symptoms, different motor manifestations were studied, including cardinal symptoms such as tremor, bradykinesia, and rigidity. In addition, other motor aspects such as FOG, dyskinesia, and gait impairment were addressed. In detail, more than half of the studies were conducted to assess gait disturbances and tremor. The most studied motor symptom was gait impairment (n = 22), which was used for computer-assisted diagnosis and prediction of disease severity. Tremor (n = 16) and FOG (n = 15) were studied to monitor disease severity. On the other hand, bradykinesia and FMI (n = 12) were evaluated for the diagnosis and monitoring of PD. Other important motor aspects such as dyskinesia (n = 3), balance (n = 1), and rigidity (n = 1) were poorly addressed. With regard to NMM, most of them focused on diagnosis and monitoring. Speech impairment (n = 11) was the most frequently assessed symptom, followed by brain dysfunction (n = 2), cognitive health (n = 2), and emotional expression disturbance (n = 1).

Wearable technology. Inertial sensors were the most common sensing technology, accounting for 67% of the studies reviewed. Specifically, the accelerometer was the most used sensor (38%), followed by the gyroscope (23%). Force sensors (14%) were used to assess gait impairment, including FOG, while the smartphone touchscreen (6%) was used to monitor bradykinesia and FMI. EEG and EMG sensors were little used (< 3%), while other types of physiological sensors were not employed (e.g., ECG, galvanic skin response). The upper body sensors were located on the wrist (18%), while the lower body sensor positions included the waist (12%), legs (23%), and feet (6%). The EMG sensor was attached to the forearm, whereas the EEG sensor was attached to the scalp. The authors developed the sensing devices in 47% of the cases, while commercially available devices were used in the remaining works. A large variety of commercial sensors have been used, including smartphones (18%), which represent the most widely used technology. Surprisingly, few studies (4%) exploited the potential of consumer smartwatches to monitor signs of PD. Instead, dedicated hardware (e.g., commercial IMUs) was mainly used.

Deep learning algorithms and inputs. CNNs were the most widely used DL model (52%), followed by RNN-based architectures (24%), ANN based on MLP (14%), and autoencoders (7%). In most studies, CNN layers were often used for feature extraction, while MLP-based blocks were cascaded to perform classification and regression tasks. CNN blocks were often fed with raw data, hand-created features, or sensor signals transformed in the frequency domain using, for example, FFT, CWT, DWT, or Mel's spectrogram. The use of CNN was more extensive than RNN, although the latter was designed to process time series and capture temporal dependencies. This could be due to the faster training time and lower computational complexity of CNN models compared to RNNs. However, the potential of RNNs to process sequential data has been exploited more than that of classical ANNs, which are usually fed with hand-created features.

In four studies, autoencoder architectures (based on MLP and CNN layers) were used to process hand-created features, often used for dimensional projection. Autoencoders were mainly used in studies related to speech disorders (Xiong & Lu, 2020; Zhang, 2017; Zhang et al., 2021), but also to assess bradykinesia (Iakovakis et al., 2020) and FOG (Mohammadian Rad et al., 2018). Only three studies proposed the combination of convolutional and recurrent layers in the same network (Dar et al., 2022; Li et al., 2020; Sigcha et al., 2020). Finally, only two studies (Oğul & Özdemir, 2022; Xia et al., 2020) reported the use of attention mechanisms in combination with recurrent architectures to improve performance in gait disorders. Attention mechanisms are often used in transformer-based architectures to analyze an input sequence

and decide which parts of the sequence are important for a specific task.

In addition, other DL approaches and variants of the standard DL architectures were proposed including recurrent CNN (Stamate et al., 2018), multitask CNN (Sigcha et al., 2021), dynamic neural network (Roy et al., 2013), GRU (Shichkina et al., 2020), BiLSTM (Lauraitis et al., 2020a, 2020b; Liu et al., 2021), probabilistic neural networks (PNN) (Oung et al., 2018), heuristic tubu optimized sequence modular neural network (HTSMNN) (AlZubi et al., 2020) and ensemble models composed of DL networks (Hssayeni, Jimenez-Shahed, Burack et al., 2021; Lauraitis et al., 2020a; Papadopoulos, Iakovakis et al., 2020; Prince et al., 2019). Surprisingly, few studies (Bikias et al., 2021; Hssayeni, Jimenez-Shahed, Burack et al., 2021; Iakovakis et al., 2020; Lin et al., 2020; Naghavi & Wade, 2022; Setiawan & Lin, 2021) used transfer learning to train models, although transfer learning is widely used in areas such as image analysis. Transfer learning is a strategy in which a source (pre-trained) model designed for a task is reused as the starting point for a secondary task.

Validation. In this review, the performance of the algorithms was not compared among the studies due to the different validation methodologies and databases used in each study. However, validation is a crucial step in the development of a model that can be generalized beyond the sample population. In some studies, DL models were trained on a subset of patients and tested on new unseen subjects. This process allows for a general model that can be used on new subjects, without the need for further training. Other studies developed subject-specific DL models, trained and tested on each (single) subject. This approach provides better performance and is more sensitive to the motor fluctuations of a single patient. However, a specific model needs to be implemented for each subject, and this increases the difficulty in model development and reduces patients' compliance. Some works used an RS k-fold CV, which does not guarantee subject independence. This increases the risk of overfitting and reduces the generalizability of the implemented models. Finally, most studies used internal validation methods (training and testing division and CV), while only two studies used external validation on a different data set (Li et al., 2020; Shalin et al., 2021).

Data. The use of proprietary or public databases varied according to the specific symptom under investigation. Publicly available data sets were frequently used for the assessment of gait (64%), speech (64%), and cognitive (60%) impairment, and for FOG detection (53%). On the other hand, public data were scarcely used for the evaluation of bradykinesia (25%). Finally, only proprietary data were used for the detection and estimation of the severity of tremor and dyskinesia. The number of subjects enrolled varies widely among studies and symptoms under investigation. The sample size was reduced in FOG detection (4 studies > 20 PD and 1 > 50) and dyskinesia (1 study > 20 PD). On the other hand, the number of subjects gradually increases, moving from the detection of tremor (9 studies > 20 and 4 > 50) to the estimation of bradykinesia (11 studies > 20 and 6 > 50) and to the analysis of speech impairment (10 studies > 20 and 7 > 50). The experimental protocol designed for data collection varied across applications. Most studies used standardized exercises and/or motor tasks (e.g., MDS-UPDRS items, walking tasks). Some studies performed data acquisition using simulated ADL in supervised settings. Finally, a limited number of works (Iakovakis et al., 2020; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; Pfister et al., 2020; San-Segundo et al., 2020) analyzed data from free-living activities carried out in unsupervised environments.

Symptom-specific algorithmic approaches. As shown in the previous sections, most studies related to the assessment of motor aspects used CNN-based architectures to process multivariate data collected with inertial sensors. In addition, RNN was the preferred method for processing data for the assessment of NMM, including speech impairment, brain dysfunction, cognitive health, and emotional expression dysfunction. For validation purposes, most of the results of these algorithmic approaches were compared with scores obtained with standardized scales such as MDS-UPDRS (Part III) or mAIMS.

More specifically, to assess gait impairment, most studies used raw signals captured by force sensors adapted in the insoles to feed CNN- and RNN-based architectures. The best performance was obtained using LSTM and BiLSTM recurrent networks. In addition, most studies used signal segmentation (i.e., sliding windows) with window sizes up to 30 s, while eight studies used entire waking recordings (i.e., 2 min) to achieve an accuracy of up to 1.0 in the diagnosis of PD and the assessment of disease severity.

For FOG evaluation, most studies used raw signals captured by accelerometers located at the ankles and thighs to feed CNN and RNN-based architectures. The best performance in FOG detection was obtained by feeding these DNNs with data transformations, including DWT and FFT. Most studies used sliding windows of a fixed size of 1–4 s, due to the need for algorithms capable of running in real time for gait assistance.

Most studies used feature sets extracted from accelerometers and gyroscopes placed on the wrist to assess tremor severity. These data were used to feed CNN, LSTM, and ANN-based architectures. Most studies reported higher accuracy using CNN and its variants (i.e. recurrent, multitask, or ensemble) in combination with sliding windows of fixed size ranging from 2 s to 5 s. No constraints on window size were identified, however, short windows were the preferred method for signal segmentation.

For the assessment of bradykinesia, most studies used feature sets extracted from data recorded by accelerometers and gyroscopes located in the upper and lower limbs. These data were commonly used to feed CNN- and ANN-based architectures to achieve higher accuracy and strong correlation. The most widely used approach was based on the use of 5–60 s sliding windows and data corresponding to the execution of MDS-UPDRS tasks (Part III). The use of larger sliding windows was found compared to other motor manifestations, probably because the symptom is assessed by observing the evolution of repetitive exercises.

All studies related to dyskinesia used accelerometers, alone or in combination with gyroscopes or sEMG sensors located in the upper and lower limbs. The feature sets and raw signals collected from these sensors were used to feed CNN, BiLSTM, and ANN-based architectures. Variable results were reported in the detection of dyskinesia and prediction of severity using scales such as mAIMS. Finally, the preferred methods for signal segmentation were 2–60 s sliding windows.

For balance assessment, a series of gait and balance characteristics were extracted from inertial sensors placed on different parts of the body. Data were processed using an ANN-based architecture, which provided an accuracy of 0.89 to discriminate between PwPD and subjects with essential tremor.

To assess rigidity, a set of features extracted from typing sessions was collected using the touch screen of a mobile phone. Data were processed with a CNN autoencoder to extract temporal information and diagnose PD. This approach reported high accuracy compared to the MDS-UPDRS assessment.

For speech disturbances, most studies have used pre-extracted feature sets obtained from speech signals captured by the smartphone microphone. These data were used to feed RNN- and CNN-based architectures. The proposed approaches were used to identify between PD and CH and estimate the severity of PD. The highest performance in PD diagnosis based on speech analysis was obtained using autoencoder-extracted feature sets, while signal segmentation information was not clearly reported.

To assess brain function, two studies used feature sets and raw data extracted from EEG and DBS systems to feed CNN- and ANN-based architectures, respectively. Sufficient accuracy in the diagnosis of PD was reported using raw EEG signals segmented into 2 s windows and a CNN, while changes in brain function were detected with high accuracy using a DBS system and a modular neural network.

For the assessment of cognitive health, two studies used memory and cognitive test data collected with a smartphone touchscreen to analyze neural and cognitive deficits. These data were used to feed ensemble architectures comprising ANN and LSTM models, in addition to classical ML (shallow) algorithms. The best results were obtained using an ensemble that included an LSTM model. In these approaches, data segmentation was not used due to the nature of the data.

Finally, emotional dysfunction was assessed in a single study. A set of features was extracted from the EEG data to detect emotional states in PwPD. Data was used to feed an architecture that combined CNN and LSTM layers. In this approach, a sliding window of 1 s was used for signal segmentation.

5. Discussion

The results of this review show that the adoption of DL techniques and wearables for PD assessment is currently at an early stage, as evidenced by the low number of published studies, compared to the number of studies employing classical ML reported in Mei et al. (2021) and Tanveer et al. (2022). However, there is a growing interest in the use of DL and wearables to assess PD, as evident from the number of studies published since 2018 (n = 66, 96%).

The main reasons for the attractiveness of DL techniques lie in their potential to provide superior performance compared to classical ML approaches and the opportunity to process data in their raw form. The latter can reduce the effort of designing feature sets limited by human knowledge in a specific task. Furthermore, DL can benefit from the rise of paradigms such as big data in healthcare, particularly for the collection of large amounts of data obtained from sources such as wearable or smart devices.

Although the combination of DL and wearable technology has been applied in areas such as diagnosis, treatment response prediction, and monitoring of motor symptoms and NMM, this review highlights several challenges that limit the implementation of these technologies in clinical practice. These challenges include the transparency of the algorithms, implementation requirements (a large amount of data and computational processing), and imbalance in the type of symptoms analyzed.

In this context, the following subsections discuss these challenges and provide recommendations for improving future work. Specifically, the discussion section is divided into the following subsections: Section 5.1 reports and discusses the challenges of PD assessment with wearables and DL algorithms. Section 5.2 provides suggestions to overcome the current challenges. Finally, the limitations of this review are summarized in Section 5.3.

5.1. Challenges

Although DL techniques have gained notoriety for producing SOTA results using large data sets, one of its main criticisms is that the resulting models are difficult to interpret. This is a major problem in the algorithmic domain, causing models to be viewed as "black boxes" due to the lack of explainability and statistical interpretability of the mechanism involved in model predictions. This situation limits the implementation of this technology in clinical practice, as the predictions could be used in decision-making to adapt patient treatments or to support disease diagnosis.

With regard to the symptoms analyzed, most of the studies reviewed (86%) focused on the analysis of motor symptoms, while the remaining studies (14%) analyzed NMM. This indicates a strong imbalance in the analysis of the two aspects, although NMM occurs in approximately 90% of patients at all stages of the disease (Chaudhuri, Odin, Antonini, & Martinez-Martin, 2011). This may be due to the fact that wearable technology is generally designed to analyze physical variables rather than cognitive aspects; this still represents a challenge in objective monitoring to provide a holistic approach to disease management. Furthermore, more than half of the studies examined in this review were conducted to assess gait disturbances and tremor. Other important motor aspects such as dyskinesia, postural instability, and rigidity were considered in less than 7% of the studies. However, these latter motor aspects have a strong impact on patient's daily life and routines, limiting motor performance and increasing the risk of falls.

For data collection, wearables have been widely used to assess motor symptoms, mainly using inertial sensors. On the other hand, a small number of wearables (e.g., microphones, smartphones, and portable EEG) have been used to monitor NMMs. This imbalance reveals the lack attention paid to monitoring NMM and the difficulty of assessing aspects related to mental health. With regard to NMM, the use of microphones and smartphones is a clear option to collect data remotely or in the clinic, given the nature of the data analyzed to assess aspects of language and cognition. However, for motor symptoms, great heterogeneity in data collection methodologies was identified, as well as low adoption of consumer devices such as smartwatches. Therefore, consensus on the number, type of sensors, and body position remains unclear, regardless of data analysis technologies.

In addition, most of the studies examined were conducted in the laboratory, often using standardized exercises (e.g., MDS-UPDRS Part III or TUG) and programmed ADLs (e.g., walking, sitting, standing, reading, and writing). Only five studies (Iakovakis et al., 2020; Papadopoulos, Iakovakis et al., 2020; Papadopoulos, Kyritsis et al., 2020; Pfister et al., 2020; San-Segundo et al., 2020) used data collected in free-living and wild-life contexts.

Often the predictive performance of the model obtained under free-living conditions tends to be lower than that obtained in the laboratory. However, results obtained under free-living conditions can better represent the symptomatology and characteristics of PD, avoiding occasional evaluations or the effect of Hawthorne observation. Therefore, there is still a need to improve the accuracy and results of systems designed for free-living monitoring, considering the complexity and heterogeneity of PD manifestations in daily life.

Furthermore, the size of the data sets used to train and evaluate the predictive models varied widely between studies. Despite larger data sets may be useful for improving the predictive power of DL-based approaches (Sarker, 2021), most of the studies reviewed in this survey considered a low number of subjects ($n < 30$). The use of data of a low number of participants may not be representative of the entire PD population, given the heterogeneity of symptoms. This situation hinders the development of generalized models that can be implemented in real-world clinical or research applications.

Finally, this review reveals that private databases were used more frequently (54%) than public databases. In specific, only private data sets were used to assess tremor and dyskinesia, whereas few public databases were used for FOG detection (14%) and the estimation of bradykinesia (25%). In contrast, public databases were used in all studies related to speech impairment and 68% of the studies related to gait impairment. According to these results, the amount of research articles to assess these latter manifestations seems to be driven by the availability of public databases. Conversely, the use of proprietary data sets and the lack of gold-standard databases can severely limit the comparison between studies, the generalizability of results, and the advancement of the SOTA.

5.2. Guidelines

To overcome the identified challenges and improve reporting standards, some guidelines are proposed in the following subsection. These could encourage the implementation of DL algorithms in combination with wearable technology as complementary tools in clinical and research contexts, as well as help improve the usefulness of current studies in future research.

First, the performance and interpretability of the model are equally important for healthcare problems. Physicians are unlikely to adopt a system they cannot understand. Therefore, model explainability and interpretability of results are key steps in developing reliable and trustworthy systems for PD evaluation (Miotto, Wang, Wang, Jiang, & Dudley, 2017). Future research on PD assessment could consider the use of interpretation tools to provide more explainable analysis methods (Li et al., 2022) to support the end-users adoption.

Second, despite the potential of combining DL and wearables for PD assessment, few studies have discussed actual implementations in clinical practice, providing limited evidence of the advantages of this technology over standard assessment methods. Therefore, clear evidence is needed to determine the benefits and validity of this technology in real-life conditions. The use of standardized tasks and supervised environments may be the basis for the development of new automatic assessment tools. Moreover, passive PD monitoring in the home environment and free-living data can help demonstrate the advantages of DL and wearables over other approaches.

Third, in regard to the area of application, most studies focused on PD monitoring (60%) and diagnosis (36%), while predicting the response to therapy and rehabilitation are underrepresented applications. These latter aspects should be further investigated because proper planning and implementation of therapy in the early stages can improve the quality of life and present a long-term benefit for the patient's health (Kilzheimer, Hentrich, Burkhardt, & Schulze-Hentrich, 2019). Furthermore, a clear imbalance in the assessment of symptoms has been observed. A large number of studies have focused on symptoms such as gait impairment, FOG, tremor, and speech disorders. In part, progress in the assessment of these symptoms is due to the existence of a wide variety of data sets, which provide a solid basis for research and the development of new methods. In this context, new open-access sources are needed to promote research on other significant features such as rigidity, dyskinesia, postural instability, and cognitive impairment.

Fourth, in regard to the evaluation of the quality report, several weaknesses were identified that can be improved in future studies. Most of the studies did not provide sufficient details on the implementation of DL models, training settings, and the validation procedure. This reduces the transparency of the reported results due to the limited evidence of accuracy in different contexts. This situation severely limits the reproducibility of results and the generalizability of algorithmic approaches to other data sets. Only a few studies (Camps et al., 2018; Deng et al., 2022; Lonini et al., 2018; O'Day et al., 2022; Papadopoulos, Iakovakis et al., 2020; Pfister et al., 2020; Pham, 2021; Prince et al., 2019; Shichkina et al., 2020; Zhao et al., 2022) published the architecture and parameters of the DL model in online repositories. In addition, a comprehensive performance evaluation was rarely reported in the studies. As can be seen in Fig. 10, accuracy was reported in less than 70% of the studies, sensitivity and specificity in less than 60%, and AUC and F–score in less than 35%. However, the AUC is one of the most important evaluation metrics to verify the performance of any classification model. Unlike sensitivity and specificity metrics, which depend on the classification threshold selected, the AUC is threshold-independent and provides an overall measure of the model's diagnostic ability. Furthermore, the F–score (i.e. the harmonic ratio between sensitivity and accuracy) is preferable to accuracy when dealing with unbalanced data sets, which are common in the assessment of PD symptoms. Regarding the regression task, correlation coefficients were always reported (Fig. 10). However, error measures were reported less frequently (MAE: 57%, RMSE: 29%), although these metrics are crucial for assessing the goodness of fit by providing the degree of deviation from target values.

Finally, making the developed algorithms and models fully and publicly available and sharing the data used for experiments is important for several reasons. First, the use of databases available online allows a direct comparison of methods and results. Furthermore, the problem of a small and unrepresentative sample can be solved by using different data sets collected under similar conditions. Finally, different datasets can be used as external validation test sets (Borzì, Sigcha, Rodríguez-Martín, & Olmo, 2023; Collins et al., 2014), which are useful for verifying the generalizability of the developed model. Likewise, the transparency of the architecture and parameters of the DL model, together with the training settings, is of paramount importance for the reproducibility of the results. Therefore, it is recommended to provide detailed information on all model-building procedures and the optimization process. Additionally, the model information should be reported in such a way that the developed algorithm can be reproduced. Finally, sharing the model and/or the code used for the experiments (as done in O'Day et al. (2022) and Roy et al. (2013)) would ensure the immediate reproducibility of the results, thus considerably reducing the time and effort devoted to their reproduction.

5.3. Limitations of this study

In this review, studies published within the last 10 years were considered, with most of them published within the last three years. The strong positive publication trend recorded is set to increase in the coming years. Therefore, a large number of new research papers will soon be available. Future systematic reviews and literature surveys will be able to make use of the larger number of available studies, providing additional information and new insights.

A total of sixty-nine studies were found that used DL and wearable sensor data to assess PD symptoms. However, these aspects were not equally represented. In particular, motor symptoms were studied much more than NMM. Furthermore, some motor aspects (e.g., gait disturbances, tremor) were addressed much more frequently than others (e.g., rigidity, dyskinesia, postural instability), with the latter being addressed in less than 4% of the articles. The same imbalance was observed in the non-motor aspects, where the number of studies on speech disorders was three times higher than the number of studies focusing on all other non-motor disorders. In summary, a strong imbalance in the assessment of symptoms was observed. Thus, comparative analyses of methods and results are limited to some aspects of PD, while others remain underrepresented or even unexplored.

Comparison of model performance between studies may provide useful information for future research studies. However, the heterogeneity of the samples under investigation and the validation methods did not allow solid conclusions to be drawn. Indeed, the use of proprietary data sets and the lack of standards for the validation phase limit the comparison of methods and results between studies.

6. Conclusions

The need for objective monitoring of PD has led to the development of systems for the automatic assessment of motor and non-motor symptoms. These systems have used wearable technology and artificial intelligence to provide objective monitoring mechanisms in various contexts, such as the laboratory, the clinic, and free life. In this context, DL methods are powerful tools that allow computers to learn from a large amount of data collected with wearable technology in a cost-effective and non-invasive manner.

According to this review, the use of wearable sensors to detect motor symptoms and NMMs is widely adopted. However, the adoption

of DL techniques as data processing mechanisms is currently under development, as evidenced by the large number of studies conducted under laboratory conditions. Therefore, the clinical adoption of these technologies is still limited in clinical practice. This is due to the lack of clear evidence of the advantages of this technology over standard data analysis methods.

Although the use of DL enabled SOTA results, this study reveals that researchers downplayed the importance of interpretability in favor of obtaining significant improvements in model performance. Furthermore, several challenges related to the transparency and interpretability of algorithms should be addressed. This would increase acceptability by physicians and support the gradual adoption of these technologies and the development of reliable and trustworthy systems. Furthermore, a significant imbalance in the area of application was observed, with monitoring being the most common area of application, and predicting response to therapy being the least studied. In addition, a small number of studies were observed on specific (but important) features of PD, such as cognitive impairment or balance. Furthermore, significant heterogeneity was identified in the sample size (most studies were conducted with data from only a few PwPD) and in the use of different algorithmic approaches. However, CNN and RNN architectures were identified as the preferred methods to analyze data collected from wearable devices, including consumer devices such as smartphones and smartwatches.

Finally, this review proposes some guidelines to overcome the identified challenges. These aim to improve reporting standards and increase the usefulness of future research by making algorithms and data available. This would allow external validation in various contexts or the application of advanced techniques such as transfer learning. Furthermore, the interpretability of the algorithms should be addressed to increase the acceptability of this technology and enable the development of reliable systems for PD assessment. Also, the application of the developed wearable solutions under real-life conditions could be useful to provide insights into the benefits of combining these technologies for the evaluation of different motor symptoms and NMMs, as well as the application of technologies such as transfer learning.

Future trends include advances in both hardware components and the algorithmic area. The former includes the use of minimally invasive wearable technologies. Consumer devices such as smartphones, smartwatches, and smart rings can be used to collect and analyze data in an environmentally friendly way. Moreover, new advances in smart clothes and smart patches would enable continuous monitoring of mobility in unsupervised contexts. Multimodal data including movement and physiological signals (e.g., ECG, PPG, GSR) would allow a better understanding of motor and non-motor aspects of PD. Regarding the development of algorithms, the collection of large amounts of data in unsupervised contexts and the use of semi-supervised DL approaches would increase the detection and estimation performance of the severity of PD symptoms, fostering the generalization capability of developed solutions.

Acronyms

PD	Parkinson's disease
NMM	Non-motor manifestations
QoL	Quality of life
MDS	Movement disorder society
UPDRS	Unified Parkinson's disease rating scale
DL	Deep learning
ML	Machine learning
SOTA	State-of-the-art
ECG	Electrocardiography
EEG	Electroencephalography
EMG	Electromyography
ANN	Artificial neural network
DNN	Deep neural network
ReLU	Rectified linear unit
CNN	Convolutional neural network
RNN	Recurrent neural network
GRU	Gated recurrent unit
LSTM	Long short-term memory
LOSO	Leave-one-subject-out
S-I	Subject-independent
S-D	Subject-dependent
RS	Random shuffle
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PwPD	Patients with Parkinson's disease
FFT	Fast Fourier transform
CWT	Continuous wavelet transform
DWT	Discrete wavelet transform
DFT	Discrete Fourier transform
PCA	Principal component analysis
MLP	Multi-layer perceptron
DAE	Deep autoencoders
MNN	Modular neural network
Bi-LSTM	Bi-directional long short-term memory
CRNN	Convolutional recurrent neural network
BFGS	Broyden–Fletcher–Goldfarb–Shanno
MCC	Matthew's correlation coefficient
MSE	Mean square error
ICC	Intra-class correlation coefficient
ADL	Activities of daily living
IMU	Inertial measurement unit
TUG	Timed-up and go
CDA	Convolutional denoising autoencoder
CV	Cross validation
GM	Geometric mean
HC	Halthy controls
VGRF	Vertical ground reaction force
RSRNA	Siamese recurrent network with attention
CorrMNN	Correlative memory neural network
FMI	Fine motor impairment
EMS	Electrical muscle stimulation
PNN	Probabilistic neural network
EWT	Empirical wavelet transform
LOTO	Leave-one-task-out
mAIMS	modified involuntary abnormal movements scale
SMOTE	Synthetic minority oversampling technique
LDA	Linear discriminant analysis
kNN	k-nearest neighbor
DBS	Deep brain simulation
SAGE	Self-administered gerocognitive examination
LR	Logistic regression
HTSMNN	Sequence-optimized modular neural network

CRediT authorship contribution statement

Luis Sigcha: Resources, Conceptualization, Supervision, Methodology, Formal analysis, Investigation, Project administration, Writing – original draft. **Luigi Borzì:** Resources, Conceptualization, Supervision, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft. **Federica Amato:** Investigation, Validation, Formal analysis, Writing – review & editing. **Irene Rechichi:** Investigation, Validation, Formal analysis, Writing – review & editing. **Carlos Ramos-Romero:** Investigation, Validation, Formal analysis, Writing – review & editing. **Andrés Cárdenas:** Investigation, Validation, Formal analysis, Writing – review & editing. **Luis Gascó:** Investigation, Validation, Supervision, Writing – review & editing. **Gabriella Olmo:** Investigation, Validation, Formal analysis, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.120541.

References

Albanese, A. (2013). Standard strategies for diagnosis and treatment of patients with newly diagnosed Parkinson disease: ITALY. *Neurology Clinical Practice*, *3*(6), 476–477. http://dx.doi.org/10.1212/01.CPJ.0000437018.37541.eb.

Alharthi, A. S., Casson, A. J., & Ozanyan, K. B. (2021). Gait spatiotemporal signal analysis for Parkinson's disease detection and severity rating. *IEEE Sensors Journal*, *21*(2), 1838–1848. http://dx.doi.org/10.1109/JSEN.2020.3018262.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. http://dx.doi.org/10.1186/s40537-021-00444-8.

AlZubi, A. A., Alarifi, A., & Al-Maitah, M. (2020). Deep brain simulation wearable IoT sensor device based Parkinson brain disorder detection using heuristic tubu optimized sequence modular neural network. *Measurement*, *161*, Article 107887. http://dx.doi.org/10.1016/j.measurement.2020.107887.

Armstrong, M. J., & Okun, M. S. (2020). Diagnosis and treatment of Parkinson disease: A review. *JAMA*, *323*(6), 548–560. http://dx.doi.org/10.1001/jama.2019.22360.

Ashfaque Mostafa, T., Soltaninejad, S., McIsaac, T. L., & Cheng, I. (2021). A comparative study of time frequency representation techniques for freeze of gait detection and prediction. *Sensors*, *21*(19), 6446. http://dx.doi.org/10.3390/s21196446.

Ashour, A. S., El-Attar, A., Dey, N., El-Kader, H. A., & Abd El-Naby, M. M. (2020). Long short term memory based patient-dependent model for FOG detection in Parkinson's disease. *Pattern Recognition Letters*, *131*, 23–29. http://dx.doi.org/10.1016/j.patrec.2019.11.036.

Bachlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J., Giladi, N., et al. (2009). Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 436–446. http://dx.doi.org/10.1109/TITB.2009.2036165.

Balaji, E., Brindha, D., Vinodh Kumar, E., & Vikrama, R. (2021). Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network. *Applied Soft Computing*, *108*, Article 107463. http://dx.doi.org/10.1016/j.asoc.2021.107463.

Balestrino, R., & Schapira, A. H. V. (2020). Parkinson disease. *European Journal of Neurology*, *27*(1), 27–42. http://dx.doi.org/10.1111/ene.14108.

Berke Erdas, C., Sumer, E., & Kibaroglu, S. (2022). CNN-based severity prediction of neurodegenerative diseases using gait data. *Digital Health*, *8*, http://dx.doi.org/10.1177/20552076221075147.

Berus, L., Klancnik, S., Brezocnik, M., & Ficko, M. (2018). Classifying Parkinson's disease based on acoustic measures using artificial neural networks. *Sensors*, *19*(1), 15. http://dx.doi.org/10.3390/s19010016.

Bhidayasiri, R., & Martinez-Martin, P. (2017). Clinical assessments in Parkinson's disease: Scales and monitoring. *International Review of Neurobiology*, *132*, 129–182. http://dx.doi.org/10.1016/bs.irn.2017.01.001.

Bikias, T., Iakovakis, D., Hadjidimitriou, S., Charisis, V., & Hadjileontiadis, L. J. (2021). DeepFoG: An IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. *Frontiers in Robotics and AI*, *8*, Article 537384. http://dx.doi.org/10.3389/frobt.2021.537384.

Bonato, P. (2010). Wearable sensors and systems. From enabling technology to clinical applications. *IEEE Engineering in Medicine and Biology Magazine*, *29*(3), 25–36. http://dx.doi.org/10.1109/MEMB.2010.936554.

Borzì, L., Sigcha, L., Rodríguez-Martín, D., & Olmo, G. (2023). Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. *Artificial Intelligence in Medicine*, *135*, Article 102459. http://dx.doi.org/10.1016/j.artmed.2022.102459.

Borzì, L., Varrecchia, M., Sibille, S., Olmo, G., Artusi, C. A., Fabbri, M., et al. (2020). Smartphone-based estimation of item 3.8 of the MDS-UPDRS-III for assessing leg agility in people with Parkinson's disease. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 140–147. http://dx.doi.org/10.1109/OJEMB.2020.2993463.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., et al. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, *3*(1), 1–9. http://dx.doi.org/10.1038/sdata.2016.11.

Camps, J., Sama, A., Martin, M., Rodriguez-Martin, D., Perez-Lopez, C., Arostegui, J., et al. (2018). Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge-Based Systems*, *139*, 119–131. http://dx.doi.org/10.1016/j.knosys.2017.10.017.

Channa, A., Popescu, N., & Ciobanu, V. (2020). Wearable solutions for patients with Parkinson's disease and neurocognitive disorder: A systematic review. *Sensors*, *20*(9), http://dx.doi.org/10.3390/s20092713.

Chaudhuri, K. R., Odin, P., Antonini, A., & Martinez-Martin, P. (2011). Parkinson's disease: the non-motor issues. *Parkinsonism & Related Disorders*, *17*(10), 717–723. http://dx.doi.org/10.1016/j.parkreldis.2011.02.018.

Chen, F., Fan, X., Li, J., Zou, M., & Huang, L. (2021). Gait analysis based Parkinson's disease auxiliary diagnosis system. *Journal of Internet Technology*, *22*(5), http://dx.doi.org/10.53106/160792642021092205005.

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., et al. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1), 1–11. http://dx.doi.org/10.1186/1471-2288-14-40.

Dar, M. N., Akram, M. U., Yuvaraj, R., Gul Khawaja, S., & Murugappan, M. (2022). EEG-based emotion charting for Parkinson's disease patients using convolutional recurrent neural networks and cross dataset learning. *Computers in Biology and Medicine*, *144*, Article 105327. http://dx.doi.org/10.1016/j.compbiomed.2022.105327.

Deb, R., An, S., Bhat, G., Shill, H., & Ogras, U. Y. (2022). A systematic survey of research trends in technology usage for Parkinson's disease. *Sensors*, *22*(15), 5491. http://dx.doi.org/10.3390/s22155491.

Del Din, S., Kirk, C., Yarnall, A. J., Rochester, L., & Hausdorff, J. M. (2021). Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: Vision, state of the art, and challenges ahead. *Journal of Parkinson's Disease*, *11*(s1), S35–S47. http://dx.doi.org/10.3233/JPD-202471.

Deng, K., Li, Y., Zhang, H., Wang, J., Albin, R. L., & Guan, Y. (2022). Heterogeneous digital biomarker integration out-performs patient self-reports in predicting Parkinson's disease. *Communications Biology*, *5*, 58. http://dx.doi.org/10.1038/s42003-022-03002-x.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387. http://dx.doi.org/10.1561/2000000039.

Dorsey, E. R., Sherer, T., Okun, M. S., & Bloem, B. R. (2018). The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's Disease*, *8*(s1), S3–S8. http://dx.doi.org/10.3233/JPD-181474.

Dunn, J., Runge, R., & Snyder, M. (2018). Wearables and the medical revolution. *Personalized Medicine*, *15*(5), 429–448. http://dx.doi.org/10.2217/pme-2018-0044.

El Maachi, I., Bilodeau, G.-A., & Bouachir, W. (2020). Deep 1D-convnet for accurate parkinson disease detection and severity prediction from gait. *Expert Systems with Applications*, *143*, Article 113075. http://dx.doi.org/10.1016/j.eswa.2019.113075.

Esfahani, A. H., Dyka, Z., Ortmann, S., & Langendörfer, P. (2021). Impact of data preparation in freezing of gait detection using feature-less recurrent neural network. *IEEE Access*, *9*, 138120–138131. http://dx.doi.org/10.1109/ACCESS.2021.3117543.

Espay, A. J., Hausdorff, J. M., Sánchez-Ferro, Á., Klucken, J., Merola, A., Bonato, P., et al. (2019). A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Movement Disorders*, *34*(5), 657–663. http://dx.doi.org/10.1002/mds.27671.

Fernandes, C., Ferreira, F., Lopes, R. L., Bicho, E., Erlhagen, W., Sousa, N., et al. (2021). Discrimination of idiopathic Parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect. *Journal of Biomechanics*, *125*, Article 110214. http://dx.doi.org/10.1016/j.jbiomech.2020.110214.

Giannakopoulou, K.-M., Roussaki, I., & Demestichas, K. (2022). Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. *Sensors*, *22*(5), http://dx.doi.org/10.3390/s22051799.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics, Vol. 9* (pp. 249–256). Chia Laguna Resort, Sardinia, Italy.

Goetz, C. G. (2011). The history of Parkinson's disease: early clinical descriptions and neurological therapies. *Cold Spring Harbor Perspectives in Medicine*, *1*(1), a008862. http://dx.doi.org/10.1101/cshperspect.a008862.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., et al. (2008). Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. http://dx.doi.org/10.1002/mds.22340.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. http://dx.doi.org/10.1161/01.cir.101.23.e215.

Han Byul, K., Woong Woo, L., Aryun, K., Hong Ji, L., Park, H. Y., Jeon, H. S., et al. (2018). Wrist sensor-based tremor severity quantification in Parkinson's disease using convolutional neural network. *Computers in Biology and Medicine*, *95*, 140–146. http://dx.doi.org/10.1016/j.compbiomed.2018.02.007.

Hausdorff, J. M., Lertratanakul, A., Cudkowicz, M. E., Peterson, A. L., Kaliton, D., & Goldberger, A. L. (2000). Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. *Journal of Applied Physiology*, http://dx.doi.org/10.1152/jappl.2000.88.6.2045.

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., et al. (2018). Wearable sensors: modalities, challenges, and prospects. *Lab on A Chip*, *18*(2), 217–248. http://dx.doi.org/10.1039/C7LC00914C.

Hssayeni, M., Jimenez-Shahed, J., & Burack, M. (2021). Dyskinesia estimation during activities of daily living using wearable motion sensors and deep recurrent networks. *Scientific Reports*, *11*, 7865. http://dx.doi.org/10.1038/s41598-021-86705-1.

Hssayeni, M. D., Jimenez-Shahed, J., Burack, M. A., & Ghoraani, B. (2019). Wearable sensors for estimation of Parkinsonian tremor severity during free body movements. *Sensors*, *19*(19), http://dx.doi.org/10.3390/s19194215.

Hssayeni, M., Jimenez-Shahed, J., Burack, M., et al. (2021). Ensemble deep model for continuous estimation of unified Parkinson's disease rating scale III. *BioMedical Engineering OnLine*, *20*, 32. http://dx.doi.org/10.1186/s12938-021-00872-w.

Iakovakis, D., Chaudhuri, K. R., Klingelhoefer, L., Bostantjopoulou, S., Katsarou, Z., Trivedi, D., et al. (2020). Screening of parkinsonian subtle fine-motor impairment from touchscreen typing via deep learning. *Scientific Reports*, *10*, 12623. http://dx.doi.org/10.1038/s41598-020-69369-1.

Ibrahim, A., Zhou, Y., Jenkins, M. E., Trejos, A. L., & Naish, M. D. (2021). Real-time voluntary motion prediction and Parkinson's tremor reduction using deep neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1413–1423. http://dx.doi.org/10.1109/TNSRE.2021.3097007.

Jane, Y. N., Nehemiah, H. K., & Arputharaj, K. (2016). A Q-backpropagated time delay neural network for diagnosing severity of gait disturbances in Parkinson's disease. *Journal of Biomedical Informatics*, *60*, 169–176. http://dx.doi.org/10.1016/j.jbi.2016.01.014.

Jankovic, J. (2005). Motor fluctuations and dyskinesias in Parkinson's disease: clinical manifestations. *Movement Disorders*, *20*(S11), S11–S16. http://dx.doi.org/10.1002/mds.20458.

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery and Psychiatry*, *79*(4), 368–376. http://dx.doi.org/10.1136/jnnp.2007.131045.

Johansson, D., Malmgren, K., & Alt Murphy, M. (2018). Wearable sensors for clinical applications in epilepsy, Parkinson's disease, and stroke: a mixed-methods systematic review. *Journal of Neurology*, *265*(8), 1740–1752. http://dx.doi.org/10.1007/s00415-018-8786-y.

Kilzheimer, A., Hentrich, T., Burkhardt, S., & Schulze-Hentrich, J. M. (2019). The challenge and opportunity to diagnose Parkinson's disease in midlife. *Frontiers in Neurology*, *10*, 1328. http://dx.doi.org/10.3389/fneur.2019.01328.

Kim, H., Lee, H., Lee, W., Kim, S., Jeon, H. S., Park, H. Y., et al. (2018). Validation of freezing-of-gait monitoring using smartphone. *Telemedicine and E-Health*, *24*(11), 899–907. http://dx.doi.org/10.1089/tmj.2017.0215.

Kobylecki, C. (2020). Update on the diagnosis and management of Parkinson's disease. *Clinical Medicine Journal*, *20*(4), 393–398. http://dx.doi.org/10.7861/clinmed.2020-0220.

Lauraitis, A., Maskeliunas, R., Damasevicius, R., & Krilavicius, T. (2020a). A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment. *Sensors*, *20*(11), 3236. http://dx.doi.org/10.3390/s20113236.

Lauraitis, A., Maskeliūnas, R., Damaševičius, R., & Krilavičius, T. (2020b). Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features. *IEEE Access*, *8*, 96162–96172. http://dx.doi.org/10.1109/ACCESS.2020.2995737.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., et al. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, *64*(12), 3197–3234. http://dx.doi.org/10.1007/s10115-022-01756-8.

Li, B., Zhiming, Y., Jianguo, W., Shaonan, W., Xianjun, Y., & Yining, S. (2020). Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors. *Electronics*, *9*(11), 1919. http://dx.doi.org/10.3390/electronics9111919.

Lin, C.-H., Wang, F.-C., Kuo, T.-Y., Huang, P.-W., Chen, S.-F., & Fu, L.-C. (2022). Early detection of Parkinson's disease by neural network models. *IEEE Access*, *10*, 19033–19044. http://dx.doi.org/10.1109/ACCESS.2022.3150774.

Lin, C.-W., Wen, T.-C., & Setiawan, F. (2020). Evaluation of vertical ground reaction forces pattern visualization in neurodegenerative diseases identification using deep learning and recurrence plot image feature extraction. *Sensors*, *20*(14), 3857. http://dx.doi.org/10.3390/s20143857.

Little, M., Mcsharry, P., Roberts, S., Costello, D., & Moroz, I. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 1. http://dx.doi.org/10.1186/1475-925X-6-23.

Liu, X., Li, W., Liu, Z., Du, F., & Zo, Q. (2021). A dual-branch model for diagnosis of Parkinson's disease based on the independent and joint features of the left and right gait. *Applied Intelligence*, *51*, 7221–7232. http://dx.doi.org/10.1007/s10489-020-02182-5.

Loh, H. W., Hong, W., Ooi, C. P., Chakraborty, S., Barua, P. D., Deo, R. C., et al. (2021). Application of deep learning models for automated identification of Parkinson's disease: A review (2011–2021). *Sensors*, *21*(21), http://dx.doi.org/10.3390/s21217034.

Lonini, L., Dai, A., Shawen, N., Simuni, T., Poon, C., Shimanovich, L., et al. (2018). Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digital Medicine*, *1*, 64. http://dx.doi.org/10.1038/s41746-018-0071-z.

Lu, R., Xu, Y., Li, X., Fan, Y., Zeng, W., Tan, Y., et al. (2020). Evaluation of wearable sensor devices in Parkinson's disease: A review of current status and future prospects. *Parkinson's Disease*, *2020*, Article 4693019. http://dx.doi.org/10.1155/2020/4693019.

Luis-Martínez, R., Monje, M. H. G., Antonini, A., Sánchez-Ferro, & Mestre, T. A. (2020). Technology enabled care: Integrating multidisciplinary care in Parkinson's disease through digital technology. *Frontiers in Neurology*, *11*, Article 575975. http://dx.doi.org/10.3389/fneur.2020.575975.

Mazilu, S., Blanke, U., Roggen, D., Tröster, G., Gazit, E., & Hausdorff, J. M. (2013). Engineers meet clinicians: Augmenting Parkinson's disease patients to gather information for gait rehabilitation. In *Proceedings of the 4th augmented human international conference* (pp. 124–127). New York, NY, USA.

Mei, J., Desrosiers, C., & Frasnelli, J. (2021). Machine learning for the diagnosis of Parkinson's disease: A review of literature. *Frontiers in Aging Neuroscience*, *13*, http://dx.doi.org/10.3389/fnagi.2021.633752.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. http://dx.doi.org/10.1093/bib/bbx044.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., & Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes*, *10*(2), http://dx.doi.org/10.3390/genes10020087.

Mohammadian Rad, N., Van Laarhoven, T., Furlanello, C., & Marchiori, E. (2018). Novelty detection using deep normative modeling for IMU-based abnormal movement monitoring in Parkinson's disease and autism spectrum disorders. *Sensors*, *18*(10), 3533. http://dx.doi.org/10.3390/s18103533.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group* (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00135.

Monje, M. H. G., Foffani, G., Obeso, J., & Sánchez-Ferro, A. (2019). New sensor and wearable technologies to aid in the diagnosis and treatment monitoring of Parkinson's disease. *Annual Review of Biomedical Engineering*, *21*, 111–143. http://dx.doi.org/10.1146/annurev-bioeng-062117-121036.

Moon, S., Song, H., Sharma, V., Lyons, K., Pahwa, R., Akinwuntan, A. E., et al. (2020). Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. *Journal of NeuroEngineering and Rehabilitation*, *17*(1), 125. http://dx.doi.org/10.1186/s12984-020-00756-5.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*, *162*(1), 1–73. http://dx.doi.org/10.7326/M14-0698.

Morgan, C., Rolinski, M., McNaney, R., Jones, B., Rochester, L., Maetzler, W., et al. (2020). Systematic review looking at the use of technology to measure free-living symptom and activity outcomes in Parkinson's disease in the home or a home-like environment. *Journal of Parkinson's Disease*, *10*(2), 429–454. http://dx.doi.org/10.3233/JPD-191781.

Mughal, H., Javed, A. R., Rizwan, M., Almadhor, A. S., & Kryvinska, N. (2022). Parkinson's disease management via wearable sensors: A systematic review. *IEEE Access*, *10*, 35219–35237. http://dx.doi.org/10.1109/ACCESS.2022.3162844.

Naghavi, N., & Wade, E. (2022). Towards real-time prediction of freezing of gait in patients with Parkinson's disease: A novel deep one-class classifier. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1726–1736. http://dx.doi.org/10.1109/JBHI.2021.3103071.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. http://dx.doi.org/10.1016/j.neucom.2021.03.091.

Noor, M. H. M., Nazir, A., Wahab, M. N. A., & Ling, J. O. Y. (2021). Detection of freezing of gait using unsupervised convolutional denoising autoencoder. *IEEE Access*, *9*, 115700–115709. http://dx.doi.org/10.1109/ACCESS.2021.3104975.

O'Day, J., Lee, M., Seagers, K., Hoffman, S., Jih-Schiff, A., Kidziński, L., et al. (2022). Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of NeuroEngineering and Rehabilitation*, *19*(1), 20. http://dx.doi.org/10.1186/s12984-022-00992-x.

Oğul, B. B., & Özdemir, S. (2022). A pairwise deep ranking model for relative assessment of Parkinson's disease patients from gait signals. *IEEE Access*, *10*, 6676–6683. http://dx.doi.org/10.1109/ACCESS.2021.3136724.

Oh, S., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., et al. (2020). A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neural Computing and Applications*, *32*, 10927–10933. http://dx.doi.org/10.1007/s00521-018-3689-5.

Oung, Q., Muthusamy, H., Basah, S., Lee, H., & Vijean, V. (2018). Empirical wavelet transform based features for classification of Parkinson's disease severity. *Journal of Medical Systems*, *42*, 29. http://dx.doi.org/10.1007/s10916-017-0877-2.

Papadopoulos, A., Iakovakis, D., Klingelhoefer, L., Bostantjopoulou, S., Chaudhuri, K. R., Kyritsis, K., et al. (2020). Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific Reports*, *10*, 21370. http://dx.doi.org/10.1038/s41598-020-78418-8.

Papadopoulos, A., Kyritsis, K., Klingelhoefer, L., Bostanjopoulou, S., Chaudhuri, K. R., & Delopoulos, A. (2020). Detecting parkinsonian tremor from IMU data collected in-the-wild using deep multiple-instance learning. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2559–2569. http://dx.doi.org/10.1109/JBHI.2019.2961748.

Park, D. J., Lee, J. W., Lee, M. J., Ahn, S. J., Kim, J., Kim, G. L., et al. (2021). Evaluation for parkinsonian bradykinesia by deep learning modeling of kinematic parameters. *Journal of Neural Transmission*, *128*, 181–189. http://dx.doi.org/10.1007/s00702-021-02301-7.

Park, Y. G., Lee, S., & Park, J. U. (2019). Recent progress in wireless sensors for wearable electronics. *Sensors*, *19*(20), http://dx.doi.org/10.3390/s19204353.

Peraza, L. R., Kinnunen, K. M., McNaney, R., Craddock, I. J., Whone, A. L., Morgan, C., et al. (2021). An automatic gait analysis pipeline for wearable sensors: A pilot study in Parkinson's disease. *Sensors*, *21*(24), 8286. http://dx.doi.org/10.3390/s21248286.

Pfister, F., Um, T., Pichler, D., Goschenhofer, J., Abedinpour, K., Lang, M., et al. (2020). High-resolution motor state detection in Parkinson's disease using convolutional neural networks. *Scientific Reports*, *10*, 5860. http://dx.doi.org/10.1038/s41598-020-61789-3.

Pham, T. (2021). Time–frequency time–space LSTM for robust classification of physiological signals. *Scientific Reports*, *11*, 6936. http://dx.doi.org/10.1038/s41598-021-86432-7.

Phokaewvarangkul, O., Vateekul, P., Wichakam, I., Anan, C., & Bhidayasiri, R. (2021). Using machine learning for predicting the best outcomes with electrical muscle stimulation for tremors in Parkinson's disease. *Frontiers in Aging Neuroscience*, *13*, Article 727654. http://dx.doi.org/10.3389/fnagi.2021.727654.

Prince, J., Andreotti, F., & De Vos, M. (2019). Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1402–1411. http://dx.doi.org/10.1109/TBME.2018.2873252.

Qin, Z., Jiang, Z., Chen, J., Hu, C., & Ma, Y. (2019). SEMG-based tremor severity evaluation for Parkinson's disease using a light-weight CNN. *IEEE Signal Processing Letters*, *26*(4), 637–641. http://dx.doi.org/10.1109/LSP.2019.2903334.

Reeve, A., Simcox, E., & Turnbull, D. (2014). Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews*, *14*, 19–30. http://dx.doi.org/10.1016/j.arr.2014.01.004.

Reich, S. G., & Savitt, J. M. (2019). Parkinson's disease. *Medical Clinics of North America*, *103*(2), 337–350. http://dx.doi.org/10.1016/j.mcna.2018.10.014.

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., & Logroscino, G. (2016). Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology*, *86*(6), 566–576. http://dx.doi.org/10.1212/WNL.0000000000002350.

Rodríguez-Martín, D., Cabestany, J., Pérez-López, C., Pie, M., Calvet, J., Samà, A., et al. (2022). A new paradigm in Parkinson's disease evaluation with wearable medical devices: A review of STAT-ONTM. *Frontiers in Neurology*, *13*, http://dx.doi.org/10.3389/fneur.2022.912343.

Rodríguez-Martín, D., Samà, A., Pérez-López, C., Català, A., Moreno Arostegui, J., Cabestany, J., et al. (2017). Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PLOS ONE*, *12*(2), http://dx.doi.org/10.1371/journal.pone.0171764.

Rovini, E., Maremmani, C., & Cavallo, F. (2017). How wearable sensors can support Parkinson's disease diagnosis and treatment: A systematic review. *Frontiers in Neuroscience*, *11*, 555. http://dx.doi.org/10.3389/fnins.2017.00555.

Roy, S. H., Cole, B. T., Gilmore, L. D., De Luca, C. J., Thomas, C. A., Saint-Hilaire, M. M., et al. (2013). High-resolution tracking of motor disorders in Parkinson's disease during unconstrained activity. *Movement Disorders*, *28*(8), 1080–1087. http://dx.doi.org/10.1002/mds.25391.

Sage Bionetworks (2016). Mpower mobile parkinson disease study. Retrieved from 10.7303/syn4993293, Accessed: December 1, 2022.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., et al. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, *17*(4), 828–834. http://dx.doi.org/10.1109/JBHI.2013.2245674.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., et al. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, *74*, 255–263. http://dx.doi.org/10.1016/j.asoc.2018.10.022.

San-Segundo, R., Navarro-Hellín, H., Torres-Sánchez, R., Hodgins, J., & De la Torre, F. (2019). Increasing robustness in the detection of freezing of gait in Parkinson's disease. *Electronics*, *8*(2), 119. http://dx.doi.org/10.3390/electronics8020119.

San-Segundo, R., Zhang, A., Cebulla, A., Panev, S., Tabor, G., Stebbins, K., et al. (2020). Parkinson's disease tremor detection in the wild using wearable accelerometers. *Sensors*, *20*(20), 5817. http://dx.doi.org/10.3390/s20205817.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(6), 1–20. http://dx.doi.org/10.1007/s42979-021-00815-1.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Setiawan, F., & Lin, C.-W. (2021). Implementation of a deep learning algorithm based on vertical ground reaction force time–frequency features for the detection and severity classification of Parkinson's disease. *Sensors*, *21*(15), 5207. http://dx.doi.org/10.3390/s21155207.

Shalin, G., Pardoel, S., Lemaire, E., Nantel, J., & Kofman, J. (2021). Prediction and detection of freezing of gait in Parkinson's disease from plantar pressure data using long short-term memory neural-networks. *Journal of NeuroEngineering and Rehabilitation*, *18*, 167. http://dx.doi.org/10.1186/s12984-021-00958-5.

Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., & Alinejad-Rokny, H. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, *113*, Article 103627. http://dx.doi.org/10.1016/j.jbi.2020.103627.

Shi, B., Tay, A., Au, W. L., Tan, D. M. L., Chia, N. S. Y., & Yen, S.-C. (2022). Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. *IEEE Transactions on Biomedical Engineering*, *69*(7), 2256–2267. http://dx.doi.org/10.1109/TBME.2022.3140258.

Shichkina, Y., Stanevich, E., & Irishina, Y. (2020). Assessment of the status of patients with Parkinson's disease using neural networks and mobile phone sensors. *Diagnostics*, *10*(4), 214. http://dx.doi.org/10.3390/diagnostics10040214.

Sica, M., Tedesco, S., Crowe, C., Kenny, L., Moore, K., Timmons, S., et al. (2021). Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLoS One*, *16*(2), Article e0246528. http://dx.doi.org/10.1371/journal.pone.0246528.

Sieberts, S. K., Schaff, J., Duda, M., Pataki, B. Á., Sun, M., Snyder, P., et al. (2021). Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's disease digital biomarker DREAM challenge. *Npj Digital Medicine*, *4*(1), 1–12. http://dx.doi.org/10.1038/s41746-021-00414-7.

Sigcha, L., Costa, N., Pavón, I., Costa, S., Arezes, P., López, J. M., et al. (2020). Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. *Sensors*, *20*, 1895. http://dx.doi.org/10.3390/s20071895.

Sigcha, L., Pavón, I., Costa, N., Costa, S., Gago, M., Arezes, P., et al. (2021). Automatic resting tremor assessment in Parkinson's disease using smartwatches and multitask convolutional neural networks. *Sensors*, *21*, 291. http://dx.doi.org/10.3390/s21010291.

Stamate, C., Magoulas, G., Kueppers, S., Nomikou, E., Daskalopoulos, I., Jha, A., et al. (2018). The cloudUPDRS app: A medical device for the clinical assessment of Parkinson's disease. *Pervasive and Mobile Computing*, *43*, 146–166. http://dx.doi.org/10.1016/j.pmcj.2017.12.005.

Steinmetzer, T., Maasch, M., Bönninger, I., & Travieso, C. M. (2019). Analysis and classification of motor dysfunctions in arm swing in Parkinson's disease. *Electronics*, *8*(12), 1471. http://dx.doi.org/10.3390/electronics8121471.

Tanveer, M., Rashid, A. H., Kumar, R., & Balasubramanian, R. (2022). Parkinson's disease diagnosis using neural networks: Survey and comprehensive evaluation. *Information Processing & Management*, *59*(3), Article 102909. http://dx.doi.org/10.1016/j.ipm.2022.102909.

Tolosa, E., Wenning, G., & Poewe, W. (2006). The diagnosis of Parkinson's disease. *The Lancet Neurology*, *5*(1), 75–86. http://dx.doi.org/10.1016/S1474-4422(05)70285-4.

Tong, L., He, J., & Peng, L. (2021). CNN-based PD hand tremor detection using inertial sensors. *IEEE Sensors Letters*, *5*(7), 1–4. http://dx.doi.org/10.1109/LSENS.2021.3074958.

Tsanas, A., Little, M., McSharry, P., & Ramig, L. (2009). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, 1. http://dx.doi.org/10.1109/TBME.2009.2036000.

van Wamelen, D. J., Sringean, J., Trivedi, D., Carroll, C. B., Schrag, A. E., Odin, P., et al. (2021). Digital health technology for non-motor symptoms in people with Parkinson's disease: Futile or future? *Parkinsonism & Related Disorders*, *89*, 186–194. http://dx.doi.org/10.1016/j.parkreldis.2021.07.032.

Varghese, J., Alen, C. M. v., Fujarski, M., Schlake, G. S., Sucker, J., Warnecke, T., et al. (2021). Sensor validation and diagnostic potential of smartwatches in movement disorders. *Sensors*, *21*(9), 3139. http://dx.doi.org/10.3390/s21093139.

Varghese, J., Fujarski, M., Hahn, T., Dugas, M., & Warnecke, T. (2020). The smart device system for movement disorders: Preliminary evaluation of diagnostic accuracy in a prospective study. *Studies in Health Technology and Informatics*, *270*, 889–893. http://dx.doi.org/10.3233/SHTI200289.

Wan, S., Liang, Y., Zhang, Y., & Guizani, M. (2018). Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. *IEEE Access*, *6*, 36825–36833. http://dx.doi.org/10.1109/ACCESS.2018.2851382.

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., et al. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One*, *15*(6), Article e0234722. http://dx.doi.org/10.1371/journal.pone.0234722.

Warmerdam, E., Hausdorff, J. M., Atrsaei, A., Zhou, Y., Mirelman, A., Aminian, K., et al. (2020). Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol*, *19*(5), 462–470. http://dx.doi.org/10.1016/S1474-4422(19)30397-7.

Wirdefeldt, K., Adami, H. O., Cole, P., Trichopoulos, D., & Mandel, J. (2011). Epidemiology and etiology of Parkinson's disease: a review of the evidence. *European Journal of Epidemiology*, *26 Suppl 1*, 1–58. http://dx.doi.org/10.1007/s10654-011-9581-6.

Xia, Y., Yao, Z., Ye, Q., & Cheng, N. (2020). A dual-modal attention-enhanced deep learning network for quantification of Parkinson's disease characteristics. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(1), 42–51. http://dx.doi.org/10.1109/TNSRE.2019.2946194.

Xiong, Y., & Lu, Y. (2020). Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. *IEEE Access*, *8*, 27821–27830. http://dx.doi.org/10.1109/ACCESS.2020.2968177.

Zhang, Y. (2017). Can a smartphone diagnose parkinson disease? A deep neural network method and telediagnosis system implementation. *Parkinson's Disease*, *2017*, http://dx.doi.org/10.1155/2017/6209703.

Zhang, X., Wang, Y., Zhang, L., Jin, B., & Zhang, H. (2021). Exploring unsupervised multivariate time series representation learning for chronic disease diagnosis. *International Journal of Data Science and Analytics*, 1. http://dx.doi.org/10.1007/s41060-021-00290-0.

Zhao, A., Li, J., Dong, J., Qi, L., Zhang, Q., Li, N., et al. (2022). Multimodal gait recognition for neurodegenerative diseases. *IEEE Transactions on Cybernetics*, *52*(9), 9439–9453. http://dx.doi.org/10.1109/TCYB.2021.3056104.

Zhao, A., Qi, L., Li, J., Dong, J., & Yu, H. (2018). A hybrid spatio-temporal model for detection and severity rating of Parkinson's disease from gait data. *Neurocomputing*, *315*, 1–8. http://dx.doi.org/10.1016/j.neucom.2018.03.032.

Zhao, N., Yang, Y., Zhang, L., Zhang, Q., Balbuena, L., Ungvari, G. S., et al. (2021). Quality of life in Parkinson's disease: A systematic review and meta-analysis of comparative studies. *CNS Neuroscience & Therapeutics*, *27*(3), 270–279. http://dx.doi.org/10.1111/cns.13549.