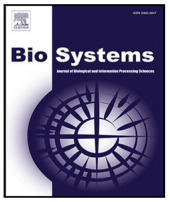
BioSystems 232 (2023) 105006

Contents lists available at ScienceDirect

BioSystems

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



Freezing of gait in Parkinson's disease: Classification using computational intelligence

Omid Mohamad Beigi [a], Lígia Reis Nóbrega [b], Sheridan Houghten [a,*], Adriano Alves Pereira [b], Adriano de Oliveira Andrade [b]

[a] Computer Science Department, Brock University, St. Catharines, Ontario, Canada
[b] Faculty of Electrical Engineering, Federal University of Uberlândia, Uberlândia, Brazil

ARTICLE INFO

Keywords:
Parkinson's disease
Freezing of gait
Classification

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disease represented by the progressive loss of dopamine producing neurons, with motor and non-motor symptoms that may be hard to distinguish from other disorders. Affecting millions of people across the world, its symptoms include bradykinesia, tremors, depression, rigidity, postural instability, cognitive decline, and falls. Furthermore, changes in gait can be used as a primary diagnosis factor. A dataset is described that records data on healthy individuals and on PD patients, including those who experience freezing of gait, in both the ON and OFF-medication states. The dataset is comprised of data for four separate tasks: voluntary stop, timed up and go, simple motor task, and dual motor and cognitive task. Seven different classifiers are applied to two problems relating to this data. The first problem is to distinguish PD patients from healthy individuals, both overall and per task. The second problem is to determine the effectiveness of medication. A thorough analysis on the classifiers and their results is performed. Overall, multilayer perceptron and decision tree provide the most consistent results.

1. Introduction

Globally, more than 8 million people are suffering from the neurodegenerative brain disorder known as Parkinson's Disease (PD). According to Dorsey et al. (2007), this number is forecast to increase drastically by the year 2030. Common symptoms of PD include bradykinesia, tremors, rigidity, postural instability, and changes in gait, including freezing of gait (Mirelman et al., 2019; Nussbaum and Ellis, 2003). Patients with PD often take shorter and slower steps than healthy individuals, although cadence is within normal parameters (Hausdorff et al., 2003).

The impairment of gait in PD patients is of increasing concern, as it affects the patient's quality of life (Hausdorff et al., 2003). Freezing of Gait (FOG) is an episodic disorder that prevents patients from starting or continuing to walk. FOG is caused by various triggers, negatively affects elevation and increases with falls, thus reducing the patient's independence (Sveinbjornsdottir, 2016; Moore et al., 2008). The lack of devices that can objectively measure this episodic gait disorder impair, for example, treatment follow-up (da Capato et al., 2015). In this sense, devices that provide more complete data, to be used in parallel with the clinical evaluation, and that provide quantitative measures, are of paramount importance as they can be used as a form of patient-centered monitoring, increasing information about the individual and the progression of the disease.

While PD symptoms may be reduced by medication, there is currently no cure (Chen, 2012). The main medication used to replace dopamine in the brain is called *Levodopa*. If the patient takes the medication regularly and correctly, then symptom fluctuations are minimized (Schaafsma et al., 2003). However, as the disease progresses, Levodopa becomes less effective, leading to an increase of motor symptoms and fluctuations. When medication is working, the patient is said to be in the ON-medication state. It is important to note that some features of gait are more affected by medication than others. In particular, a controlled study (Morris et al., 2001) showed that peak gait speed and stride length were sensitive to Levodopa, while temporal variables such as cadence, duration of swing and stance phases were resistant to the medication (Andrade, 2017).

This paper describes the creation of a dataset focused on FOG. Inertial sensors are used to collect data from both healthy individuals and from patients with PD (ON and OFF-medication states) while performing tasks that are triggers for FOG, such as gait initiation, passing through a narrow passage, left and right 180-degree and 360-degree turns to contour obstacles, 180-degree turns in place, and reaching a target. Because FOG usually happens during the OFF-medication state (Okuma, 2014b,a), it is important to carry out the experiment in both states.

* Corresponding author.
E-mail address: shoughten@brocku.ca (S. Houghten).

https://doi.org/10.1016/j.biosystems.2023.105006
Received 31 March 2023; Received in revised form 20 August 2023; Accepted 20 August 2023
Available online 25 August 2023

Utilizing the above-mentioned dataset, we assess a set of classifiers in addressing two distinct problems. The first problem involves categorizing participants as either having PD or not. The second problem focuses on differentiating between patients with PD and FOG in the OFF and ON-medication states, i.e. assessing the effectiveness of the medication. To provide a comprehensive evaluation of the classifiers, the study incorporates a detailed analysis using confusion matrices, partial dependence plots, and calibration curves. The addition of these techniques enables a thorough understanding of the classifiers' performance, capturing various aspects such as classification accuracy, dependency on input variables, and probabilistic prediction calibration. This in-depth analysis ultimately supports a more robust assessment of the classifiers' suitability for the given problems.

The remainder of this paper is organized as follows. Section 2 presents background information and discusses related work. Section 3 describes the creation of the dataset. Our methodology is described in Section 4, with the results presented and discussed in Section 5. Section 6 presents conclusions and discusses possible future work.

2. Background and related work

The gait of an individual, including those affected by PD, can be captured by various sensors attached to different locations on the body of the patient. Examples include sensors under the foot as in Yogev et al. (2005), wearable technology as in Ricci et al. (2019), and smartphones as in Ellis et al. (2015). The exact setup employed to construct the dataset for the current study is described in detail in Section 3.

In 2012, an editorial note (Chen, 2012) showed that there are many unanswered questions about the ON-medication state of FOG. The author says that the real incidence and prevalence of this condition are not well known, but probably are under-reported and not much discussed. It concludes that more studies are needed to characterize the phenomenon of FOG and the type of patients in whom it occurs. This provides motivation for the creation of datasets, such as the one we present, to further evaluate this aspect.

The different clinical features of FOG in the ON-medication state compared to continuous gait symptoms in PD can be useful in identifying patients with this phenomenon (Chen, 2012). In Schaafsma et al. (2003) nineteen patients with PD and a history of FOG, i.e., clinically significant FOG during the OFF-medication state, were recorded while they walked for 130 metres. The experiment was performed during the ON and OFF-medication states so that the gait of these patients could be further analyzed, and the researchers could evaluate the effect of Levodopa on different types of FOG. The results suggest that Levodopa raises the threshold for FOG to occur but does not cure the episode.

In Spildooren et al. (2010) participants were asked not to take their first dose of medication in the morning, to carry out the experiment, in which patients walked for 5 metres and turned left or right around a marker on the floor. In this study, it was found that the trigger that most caused the FOG episode was to flip 360 degrees in dual tasking. Three groups were assessed: healthy individuals, PD patients who do not experience FOG (*FOG-*) and PD patients who do experience FOG (*FOG+*). During the turn, the healthy and FOG- individuals decreased the cadence, while in the FOG+ individuals this increased. In addition to these, other studies, such as Hausdorff et al. (2003), Bartels et al. (2003), and Nieuwboer et al. (2001) performed experiments with the patients in both states.

Upon processing PD gait data, various algorithms can be applied to the resulting features in order to detect whether or not an individual has PD and to assess other aspects of their gait. These same algorithms can also be applied to other PD data, such as that related to speech or handwriting. We categorize these techniques into three main groups, namely, Traditional Algorithms, Evolutionary Computation, and Artificial Neural Networks, and briefly discuss related work in each of these categories.

2.1. Traditional algorithms

Having advantages such as fast implementation, producing valid results with the least amount of data, and requiring minimum features, traditional algorithms have been used by researchers over decades, including to analyze gait. For example, in Balaji et al. (2020) Support Vector Machines (SVM), were used as a supervised technique to detect the severity of PD by analyzing statistical and kinematic gait features of patients. In the mentioned study, two tasks of PD classification were considered. The first task classified PD diagnosis based on their gait behavior, while the second task rated the severity of the disease. The data was obtained by asking volunteers to wear 16 sensors. The collected data from these sensors consisted of time-series based records from different parts of the patients' feet.

In Shetty and Rao (2016), a Gaussian RBF based kernel with an SVM was applied to distinguish PD from other neuro-degenerative disorders. The focus of the study was on the particular gait features that would help to distinguish PD from other neurological disorders such as amyotrophic lateral sclerosis (ALS) and Huntington's disease, as well as healthy controls. A number of statistical feature vectors were assessed by making use of temporal gait feature records, which were then condensed by using a correlation matrix. The best feature vectors were fed to a SVM that took advantage of a Gaussian radial as the kernel.

In Elkurdi et al. (2018) K-Nearest Neighbor (KNN) was applied to the kinematic features of gait, which led to an 86.19% accuracy. Using the Microsoft Kinect v2 camera, this research investigated linear kinematic gait analysis. This analysis was based on the skeletal positioning data of the joints in the lower body (hips, knees, and ankles). Participants who walked in front of the camera at one of three distinct speeds provided the quantitative data that was gathered (slow, normal, and fast speed walks).

In Moon et al. (2020), researchers applied various classifiers to identify essential tremors for PD. The study used wearable sensors to collect the data, and demonstrated the effectiveness of the methods with a variety of measurement metrics, such as accuracy, recall, precision, and F1 score. Balance and gait characteristics features were gathered from individuals with PD and those with epilepsy during an instrumented stand and walk test. Using F1-scores, the performance of a number of different machine learning methods, such as neural networks, KNN, Decision Tree, SVM, Random Forests, and Gradient Boosting, were compared with Logistic Regression and Naive Bayes.

In Drotár et al. (2014), feature extraction and SVM were used on hand-writing data in diagnosis of motor disorders of neurodegenerative diseases, resulting in 85.6% accuracy. The study presented the PaHaW dataset which was generated by allowing patients to perform eight different handwriting tasks. Furthermore, later work in Drotár et al. (2016) thoroughly examined new pressure and stroke features linked to handwriting dynamics. Afterwards, to demonstrate 81% accuracy, three distinct classifiers were used: KNN, AdaBoost ensemble, and SVM.

For remote tracking of PD progression, in Eskidere et al. (2012) applications of machine learning frameworks SVM, LSSVM, MLPNN, and general regression NN were described when applied to the analysis of PD speech patterns. The research was based on a telemonitoring dataset that included speech recordings from early-stage PD patients with 26 features. The model worked with non-linear voice features, both classical and non-classical.

2.2. Evolutionary computation

Genetic Programming (GP) has been applied several times to the analysis of gait of PD patients. Testing against a medium-size dataset consisting of pressure sensors on each foot, in Hughes et al. (2019a) GP was found to perform and generalize better than most of the traditional algorithms, such as linear regression, when applied to this task. This work applied features from pressure sensors to a GP model in

order to introduce a descriptive symbolic nonlinear model. Later work compared GP to Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANN) (Hughes et al., 2019b), and also incorporated analysis of data of patients walking while under cognitive load (Hughes et al., 2020).

Furthermore, GP was used in Hughes et al. (2016) to develop models of human gait using data from patients' smartphones. GP was used to construct an identity gait fingerprint for two individuals, whose walking data was acquired from the accelerometer in a commercially accessible phone. Users walked freely without a predetermined routine at a regular, nonuniform speed while the phone was freely in their pocket. This data gathering process corresponds more directly to the uses of such technology in the real world. GP can also be combined with various methods to overcome overfitting.

2.3. Artificial Neural Networks (ANN)

The state-of-the-art methods of ANNs have shown a great improvement in various types of classifications. For big datasets with a large dimension of features, ANNs such as Convolutional Neural Networks (CNNs) can best handle the high dimensionality, and extract features from them. Furthermore, they have shown a good semantic interpretation by avoiding complex feature engineering.

In Hughes et al. (2019b), a feed-forward Multilayer Perceptron was used to demonstrate the effectiveness of the ANNs in terms of feature extraction and gait classification. Using more advanced algorithms, Salgueiro et al. (2021) developed a Deep Reinforcement Learning Algorithm to detect PD using data gathered from smartphones. Using Long Short-Term Memory (LSTM) networks, the authors of Oktay and Kocer (2020) differentiated both essential tremors and PD. For this matter, they used a Leap Motion Controller (LMC) which collected 3D data for classification. The tests demonstrated 90% accuracy in combined tremor analysis. In Hathaliya et al. (2022) CNNs were used to classify PD and healthy people for an accurate diagnosis of essential tremors and PD tremors. In this study, accelerometer sensors gathered patient data to measure the right and left-hand tremor data in 3 axes, which was later normalized. The model was trained on the normalized data, producing 92% accuracy.

A novel AI-based technique was created in Parisi et al. (2018) to aid in the early identification of PD. The UCI Machine Learning database was used to get dysphonic measures and clinical points in 68 patients. Feature selection was based on weights acquired from the Multi-Layer Perceptron (MLP). This reduced collection of characteristics was then used as input for the Lagrange Support Vector Machine (LSVM) classification procedure. In Sivaranjini and Sujatha (2020), MR images of healthy and Parkinson's patients were analyzed using deep neural networks. The CNN architecture AlexNet was used to improve PD diagnosis. Classification was performed after it was retrained using a pre-trained deep network using MR images. An accuracy value of 88.9% was reached. Moreover, in Oktay and Kocer (2020) a tremor classification model based on LSTM was proposed. The authors combined LSTM with CNN and attained a 90% accuracy rate. The difficulty with the above works is that they did not identify the patient's risk depending on the intensity of the tremor, and did not track illness progression depending on tremor severity. The course of illness assists patients in identifying their risk level early on and implementing preventative actions.

3. Data collection

In this section we describe the collection of data, which will then be used for various classification tasks in Section 4. The system used to collect data has been validated in López-Blanco et al. (2018). To avoid any issues associated with data loss, all collected signals were visually examined to ensure the absence of discontinuities and undesirable trends.

3.1. Methods

This study was approved by the local ethical committee[1] and conducted at the Federal University of Uberlândia, Brazil. A detailed explanation regarding the experiment was given to the participants prior to enrollment and all subjects gave their written consent to participation in the study.

3.2. Groups

Three groups were formed, the GC (Control Group) with 10 healthy volunteers, the GFOG- group, formed by 10 subjects diagnosed with PD who do not have FOG, and the GFOG+ group, also with 10 subjects, formed by those PD patients with episodes of freezing.

For our study, the total of thirty subjects were enrolled if they met the eligibility criteria and signed an informed consent prior to the study. Exclusion criteria were if the participants present severe visual and auditory impairments, have other musculoskeletal or neurodegenerative diseases, or use medications that can cause vertigo or imbalance.

There are six female and four male participants in each group. The PD patients had idiopathic PD and were able to walk unassisted in the OFF-medication period. The subjects completed several questionnaires to measure their clinical characteristics. The New Freezing of Gait Questionnaire (New FOG-Q) was used as a subjective measure of freezing of gait severity (Wang et al., 2020; Nieuwboer et al., 2009). The Movement Disorder Society – Unified Parkinson's Disease Scale (MDS-UPDRS) part II and III measured the daily living activities (Part II) and the severity of motor symptoms in PD (Part III) (Goetz et al., 2008). The mini mental state examination (MMSE) assessed cognitive functioning. The GFOG+ participants (60.8 ± 7.48 years old) have been diagnosed with PD for 12.5 ± 5.75 years, their MDS-UPDRS Part II score is 17.1 ± 10.78, with mean of 1.5 for item 13, "Freezing of gait". MDS-UPDRS Part III score in OFF-medication state is 57.5 ± 27.66, and MDS-UPDRS Part III score in ON-medication state is 41.7 ± 24.47. New FOG Questionnaire (New FOG-Q) is 19.2 ± 4.7 and MMSE score is 25.5 ± 3.92. The GFOG- participants (65.1 ± 4.38 years old) have been diagnosed with PD for 6.9 ± 4.58 years, MDS-UPDRS Part II score is 9.9 ± 3.75, MDS-UPDRS Part III score in OFF-medication state is 41.2 ± 15.51, and MSD-UPDRS Part III score in ON-medication state is 29.7 ± 14.27. MMSE score is 26.6 ± 2.27. The GC participants (64.7 ± 6.76 years old) completed only the MMSE (26.7 ± 2.21) test before the experiment.

Using the Shapiro–Wilk normality test, the distribution of the participants age are not significantly different from normal distribution, from the output (p-value > 0.05) of GFOG+ (p-value = 0.8674), GFOG- (p-value = 0.3105), and GC (p-value = 0.3998), we can assume their normality.

3.3. Technology

For data collection, inertial sensors used to collect gait and episode information of FOG are coupled to three smartwatches, as shown in Fig. 1, with a Movement Disorders Monitoring System (NetMD), which was developed by the Spanish research group CAR-CSIC in order to analyze and remotely and continuously monitor movement disturbances through inertial signals (Cabral and Andrade, 2020). This Movement Disorder Monitoring System (NetMD) is based on the joint action of an Android mobile device with smartwatch devices (Smartwatch3 SWR50 model, from Sony), with communication being established via Bluetooth. Through this system, it is possible to acquire signals from the internal gyroscopes and accelerometers of smartwatches with a frequency of sampling rate of 50 Hz (that is, with a temporal resolution of

[1] CAAE: 38885720.3.0000.5152.

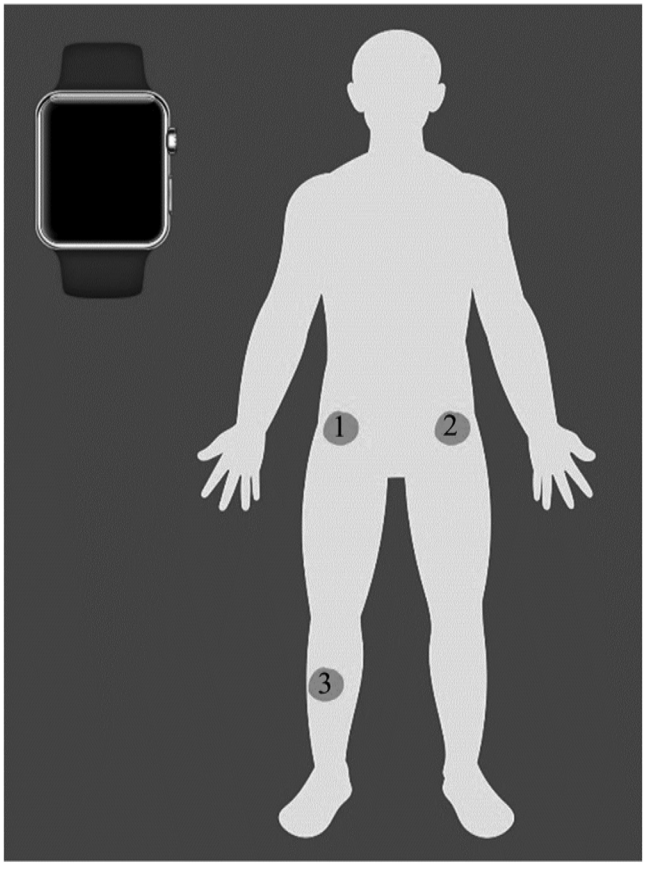


Fig. 1. Configuration of smartwatch and sensors.

20 ms). The system generates a text file with 10 columns containing the values of the inertial signals for each smartwatch (time in milliseconds, sensor name, battery status, and accelerometer and gyroscope on the x, *y* and z axes) and this file is stored in the Sony Android mobile device. After performing a time vector interpolation, it is possible to resample the data with a rate of 100 Hz, temporal resolution of 10 ms.

3.4. Sensor position

In the present work, 3 wireless inertial sensors were used, two arranged in the pelvis (Andrade et al., 2017) over the two ends of the iliac spine and one over the fibula (Moore et al., 2008; Mileti et al., 2017; Saad et al., 2014). The peroneal muscle sensor was placed on the side most affected by PD, identified by skilled and experienced healthcare professionals.

3.5. Tasks

The study makes a comparison between groups as they perform four tasks. Voluntary Stop; TUG: Timed Up and Go; Simple motor task; Dual motor and cognitive task. An image of a sample time series for these four tasks may be viewed at the following location: https://doi.org/10.5281/zenodo.6800545.

3.5.1. Voluntary stop

This task consists of the individual getting up from a chair, walking for 3 m and standing inside a square of tape made on the floor for 10 s, then returning along the same path and sitting down. This task is necessary so that the accelerometer and gyroscope signal collected during voluntary arrest can later be compared with the involuntary stop, which is freezing of gait (FOG). It was performed in the same environment as Task 2, TUG, presented below.

3.5.2. Timed Up and Go (TUG)

This test is a clinical assessment that emerged in 1991 (Podsiadlo and Richardson, 1991) and consists of recording the time required for the volunteer to get up from a chair, walk 3 m, turn (u-turn), return the same way and sit down again (Sijobert et al., 2017). It has been used in several studies addressing FOG (Sijobert et al., 2017; Nilsson and Hagell, 2009; Almeida and Lebold, 2010; Ly et al., 2017; Zirek et al., 2018).

3.5.3. Simple motor task

This is a motor circuit, in which the patient performs routine actions such as sitting and rising from a chair and walking, as well as movements that can trigger FOG, such as passing through narrow passages (door) and turning right and left around obstacles (cones on the ground). The person gets up from the chair, walks for 3 m and passes through a narrow passage of 67.5 cm. They then walk 1.30 m and go around the two obstacles, making a path in the shape of ∞. Initially, the volunteer goes around the obstacle on their worst side (identified by the physical therapist during the clinical evaluation) performing a 360° turn, then goes around the second obstacle, also with a 360° turn. Finally, the volunteer returns to the first obstacle making only one 180° turn. The volunteer returns to the chair (passing through the narrow passage again) and sits down.

3.5.4. Dual motor and cognitive task

In this stage of the experiment, the patient performs the same Simple Motor Task and, simultaneously, performs a cognitive task. The cognitive task is the Digit-Monitoring Task (DMT). DMT consists of assigning a one-digit number to each volunteer. This volunteer is instructed to count silently (without the help of the fingers) the number of times this digit is announced in the audio. The audio, which is the same for all volunteers, was transcribed so that the frequency of each digit is known by the researcher, and there is a draw of the digit before collection so that the test is randomized (Beck et al., 2015). At the end of the experiment, the participants inform the researcher the number of times they heard the digit [14]. The auditory inter-stimulus interval is displayed randomly, ranging from 100 ms to 1000 ms, to avoid gait synchronization with the audio track. The audio lasts for 60 s, and the participants were instructed to continue counting the digit even if they complete the circuit before the audio ends (Beck et al., 2015).

3.6. Medicine status - on/off

The OFF period occurs before the time determined for the next dose of the medication, in which the patient feels that the effect of the medication has worn off. When the patient is under the influence of drugs and feeling better, the period is classified as ON.

When the patient is in the OFF period, their gait is characterized by reduced or absent arm swing, reduced trunk rotation, forward leaning of the trunk, reduced range of motion of the hip, knee and ankle, slowness, reduction in step amplitude and decreased foot displacement height during the swing phase (Andrade, 2017). Another aspect noted is the increase in the contact time of the foot with the ground, called the double-stance phase.

In order for the PD volunteers to be in the OFF period, they were asked not to take the first dose of the day, staying 12 h without the medication Levodopa (Bartels et al., 2003; Wang et al., 2020).

4. Methodology

Using the data described in Section 3, the main objective is now to perform classification related to two research problems. The first problem is to distinguish between a healthy person and a patient with PD. The second problem is to determine the effectiveness of medication on each individual patient.

As a first step, the gathered data from Section 3 is preprocessed and cleaned. Furthermore, the data records are passed to the feature extraction module. Subsequently, multiple classifiers are trained on the extracted features, cross validated, and tested against the main dataset. Each of the above-mentioned steps are elaborated upon in the following subsections.

4.1. Preprocessing and feature extraction

The data records received from sensors through Bluetooth might experience a slight disconnection, which leads to a loss of data record in the dataset. Moreover, due to the time-series nature of the dataset, the length of the received signals may vary from each other by a few milliseconds. Although the validity of the data records was verified by human visualization, the structure of the data required justification. Hence, as an initial preprocessing stage of the framework, individuals with missing sensors were omitted, those with extra records were trimmed to the average record length, and sets with a few missing records were then padded. The padded sets are those with all the sensors but missing a few milliseconds (which is still processable). However, in the case of missing sensors, it is not possible to make a classification or prediction, and hence they need to be excluded from the classification process.

The records were then passed to the feature extraction module. In this module, over 200 features were extracted to help better identify the characteristics of each sensor at a particular time. To help extract these features, the Tsfresh[2] library was used. The complete list of features used in this study is referenced under "overview of extracted features" in Tsfresh. These include, for example, absolute energy, absolute maximum, and absolute sum of changes.

As the final preprocessing step, all the extracted features were stored in CSV files. At this step, all the *Not A Number* (NaN) records were removed.

4.2. Cross validation

As an integral part of the classification process, a 5-fold cross validation technique was introduced to avoid overfitting, and also to ensure the consistency of the accuracy results. In this regard, the training dataset is divided into 5 distinct subsets, and these are trained and tested one subset at a time. These are cross validated against each other; at each iteration, 4 of the 5 subsets are used as training data and the remaining one used as testing data, with each subset used as testing in one iteration. The mean results are then reported.

4.3. Classification algorithms

4.3.1. K-Nearest Neighbor (KNN)

As one of the top 10 algorithms in data mining, KNN provides an effective method for classification (Wu et al., 2007). This study takes advantage of the extracted features from the first phase, and by considering the cross validation method, they are fed to the KNN input layers using 2 neighbors.

4.3.2. Decision Tree (DT)

DT is a simple yet effective supervised machine learning algorithm that formulates the model in the form of a tree structure. Following this, the algorithm iteratively determines the nonlinear connection between the input and output by dividing the dataset into smaller subsets in the form of nodes and branches. In this study, to measure the quality of split, gini function is used.

Ensemble techniques integrate many decision tree classifiers that learn the same target function to improve classification performance by combining the predictions from each classifier. One of the main benefits of this is that combining the results of numerous classifiers reduces the probability of a bad classifier.

[2] https://tsfresh.readthedocs.io/

4.3.3. Linear Support Vector Machine (LSVM)

Introduced in Boser et al. (1992), SVM demonstrates a good fit for binary classification in a supervised manner. As an advantage, SVM deals with both linear and non-linear classifications. In this regard, SVM uses the basic notion of hyperplane and the margin to translate nonlinear inputs to higher dimensional feature space using kernel algorithms. This study used a linear kernel, and regularization parameter of 1.

4.3.4. Artificial Neural Network (ANN)

As a state-of-the-art algorithm, ANNs are defined as an interconnected groups of nodes that are used in a supervised manner. Multilayer Perceptron (MLP) is the simplest type of ANN, in which all the nodes in a layer are connected, and it provides a classification probability based on feedforward result. For the sake of simplicity, this research takes advantage of MLP to evaluate the effectiveness of ANNs in comparison to traditional methods. In this regard, only two hidden layers with "relu" activation layer are used. The weight optimization is done with Adam optimizer, and the L2 regularization term is set to 0.0001.

4.3.5. Random Forest (RF)

Random forest is an ensemble learning approach for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average forecast of the individual trees is returned for regression tasks. To classify PD patients and healthy volunteers, this study took the features from preprocessed data and trained the RF based on it. In this study, 100 trees were used in the forest with quality measurement function of gini for splitting the data.

4.3.6. Bayes Classifier (BC)

The Bayes classifier, based on the Bayesian theorem, is a statistical classifier that conducts probabilistic feature prediction and utilizes the forecast to categorize a new test dataset. When the input dimensionality is high and the prior probabilistic model is known, BC is an excellent choice. The classes' information is available in Khoury et al. (2019). Probabilistic connection modeling BC uses posterior probabilities to determine the class using the feature vectors and class variables.

4.3.7. Quadratic discriminant analysis (QDA)

QDA is a classification method that constructs class probability distributions for a given dataset by leveraging Bayes' rule and adopting a quadratic decision boundary. Discriminant analysis (DA) encompasses approaches that serve the dual purpose of reducing dimensionality and classifying instances. In addition, QDA's appeal stems from its capability to handle non-linear data by virtue of its DA-inspired framework. The underlying assumption of QDA is that each class follows a Gaussian distribution.

5. Results

Each of the algorithms specified in Section 4.3 were applied to the two research problems, namely, to distinguish between a healthy individuals and those with PD, and to determine the effectiveness of medication. Accuracy, F1-score, recall and precision are used as evaluation metrics.

In each of the following, *True Positives* (*TP*) are correct positive predictions, while *False Positives* (*FP*) are incorrect positive predictions. *False Negatives* (*FN*) denote positive labels incorrectly classified as negative, and *True Negatives* (*TN*) are correctly predicted negative labels.

Accuracy is defined as the ratio of correctly classified volunteers to the total number of volunteers:

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

Table 1
Accuracy, F1 score, precision, and recall for PD patients vs healthy volunteers on all tasks.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	86%	87%	90%	84%
Decision Tree	96%	94%	94%	99%
Linear SVM	94%	95%	93%	96%
Random Forest	76%	75%	87%	65%
Naive Bayes	74%	74%	81%	68%
Multilayer Perceptron	96%	96%	94%	99%
QDA	61%	65%	62%	48%

Table 2
Accuracy, F1 score, precision, and recall - PD patients vs healthy volunteers - Task: Voluntary stop.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	72%	66%	100%	49%
Decision Tree	100%	100%	100%	100%
Linear Support Vector Machine	72%	72%	80%	66%
Random Forest	63%	60%	75%	49%
Naive Bayes	60%	60%	75%	49%
Multilayer Perceptron	63%	66%	66%	66%
QDA	63%	66%	66%	66%

Precision is the ratio of correct positive predictions to the total positive predictions:

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

Recall is the proportion of true positive predictions relative to the sum of true positives and false negatives:

$$Recall = \frac{TP}{TP + FN}$$

Finally, the *F1-score* is the weighted harmonic mean of precision and recall:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.1. Distinguishing PD patients from healthy individuals

For this problem, all of the data from all of the tasks described in Section 3.5 are utilized. Only healthy volunteers and PD patients who are FOG+ and in the ON-medication state are taken into consideration. Therefore, the purpose is to differentiate between these two categories.

The results presented in Tables 1–5 were obtained from 9 individuals from each of these groups. Each individual carried out the same task three times at different time intervals to ensure result validity. This results in 27 healthy subjects and 27 PD subjects for each task, each comprising 200 features. In this experiment, the initial 67% of the data was employed for training, while the remaining 33% was designated for validation. The training dataset was further divided into five distinct subsets for cross-validation purposes. During each iteration, four of these subsets served as the training data while the fifth was treated as testing data. This process ensured that every subset was used as testing data once.

Table 1 shows the results from the combination of all tasks. When combining all tasks together a total of 108 (i.e. $27 * 4$) subjects were categorized as healthy, and another 108 subjects were classified as PD. The findings point to a challenge for several of the traditional algorithms, in which the models' high dimensional features act as a bottleneck. These algorithms include Random Forest and Bayes Classifier, and the results of both are quite similar. It is evident that while the majority of the classifiers display satisfactory results, the MLP outperforms others, achieving a 96% F1-score, 96% accuracy, and 99% recall. The ANN model discriminates feature dimensions more effectively compared to prior methods, particularly when considering the multitude of features extracted from the data.

Tables 2–5 provide the results when considering each of the tasks separately. As can be seen in Table 2, Decision Tree depicted a decent classification output on the Voluntary Stop. Since the Voluntary Stop task is the most simple task compared to the others, patients demonstrated very similar behavior to the healthy volunteers, which caused the remaining classifiers to struggle with detection.

It is clear to see that, in general, the Timed Up and Go (TUG) task, which is shown in Table 3, yielded the best possible outcomes for the participants. The Decision Tree, KNN, LSVM, and MLP classifiers each obtained a score of 100% across the board for every metric. Therefore, using any of these classifiers to differentiate between people with PD and healthy ones ought to provide excellent results for a TUG test.

Regarding the basic motor task that is summarized in Table 4, all but one of the classifiers achieved a precision of 100%, which means that they were able to accurately detect true positives each and every time. Notably, the Decision Tree obtained over 90% in all metrics.

Some classifiers struggled with the task that required both motor and cognitive ability, as seen in Table 5. It is probable that the data for this task have higher variability, which might cause problems for some of the classifiers. Having said that, it is important to highlight the fact that four of the classifiers still had an accuracy and F1-score of at least 90%, and five of them had a recall of 100%.

5.2. Effectiveness of medication

This experiment considers data from PD patients who are FOG+ and in the ON state for medication, along with data from PD patients who are FOG+ and in the OFF state for medication, as described in Section 3.6. As such, each of these groups comprises a total of 108 (i.e. $27 * 4$) subjects. For each subject, 200 features were extracted from their movement. In this study, the first 70% of the data was designated for training and the remaining 30% was allocated for validation. Furthermore, the training dataset was split into five subsets for cross-validation. As in the previous task, each subset served as testing data once during the five iterations.

The classifiers utilized for this experiment are the same ones that were used in the previous experiment, and the results are summarized in Table 6. Because it has a score of 99% in all four categories of accuracy, F1-score, precision, and recall, the Decision Tree seems to be the most effective method of categorization for this issue. However, a number of other methods, most notably MLP and KNN, also produced very good outcomes, with scores above 90% for every metric, in most of the tasks. Each of these (KNN, DT, MLP) could be used with a high degree of confidence to assess the effectiveness of medications.

5.3. Confusion matrix

In Fig. 2, four of the most noteworthy confusion matrices relating to single-trial classification are depicted. These were selected based on the highest counts of true negatives (TNs) and true positives (TPs) in

Table 3
Accuracy, F1 score, precision, and recall - PD patients vs healthy volunteers - Task: TUG.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	100%	100%	100%	100%
Decision Tree	100%	100%	100%	100%
Linear Support Vector Machine	100%	100%	100%	100%
Random Forest	81%	82%	87%	77%
Naive Bayes	56%	69%	57%	%88
Multilayer Perceptron	100%	100%	100%	100%
QDA	58%	63%	54%	75%

Table 4
Accuracy, F1 score, precision, and recall - PD patients vs healthy volunteers - Task: Simple motor task.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	50%	33%	100%	19%
Decision Tree	93%	94%	100%	90%
Linear Support Vector Machine	81%	82%	100%	70%
Random Forest	68%	66%	100%	49%
Naive Bayes	68%	66%	100%	49%
Multilayer Perceptron	75%	74%	100%	59%
QDA	50%	60%	60%	60%

Table 5
Accuracy, F1 score, precision, and recall - PD patients vs healthy volunteers - Task: Dual motor and cognitive task.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	93%	90%	83%	100%
Decision Tree	93%	90%	83%	100%
Linear Support Vector Machine	93%	90%	83%	100%
Random Forest	68%	61%	50%	80%
Naive Bayes	56%	58%	41%	100%
Multilayer Perceptron	93%	90%	83%	100%
QDA	87%	74%	100%	59%

Table 6
Accuracy, F1 score, precision, and recall - Effectiveness of medicine on patients.

Classifier	Accuracy	F1-Score	Precision	Recall
K-Nearest Neighbors	95%	95%	99%	92%
Decision Tree	99%	99%	99%	99%
Linear Support Vector	87%	88%	80%	100%
Random Forest	57%	58%	55%	62%
Naive Bayes	61%	59%	60%	58%
Multilayer Perceptron	96%	95%	92%	99%
QDA	55%	60%	53%	70%

Tables 1–5. Additionally, the confusion matrix is utilized on test cases (specifically, as the forecast data is required for analysis). It is apparent that the Decision Tree Classifier outperformed the other classifiers, showcasing its efficacy in achieving the highest accuracy and F1 score. QDA comes in second, with five correct predictions. In third place, the Neural Network displayed a stronger inclination towards the "False" class, indicating a lack of data for this classifier type. Lastly, the Linear SVM produced the most varied labeling and accurately predicted half of the labels.

Fig. 3 illustrates the confusion matrix for all trials combined. As seen in Fig. 3, more TPs and TNs are positioned. Consequently, having more data (additional features) guides the classifiers in the right direction. As mentioned earlier, the chart displays the number of test cases corresponding to the number of patients. Nevertheless, as outlined in Section 3, not all volunteers reached the feature extraction stage due to insufficient sensors, inadequate data, etc. In this context, the Decision Tree has the highest counts of TPs and TNs, with no incorrect predictions. It is clear that the data leans more towards the "True" class rather than the "False" class. With an increased amount of training data, the Neural Network ranks second, with 15 correctly predicted labels. Although the Linear SVM performed the poorest in the single trial, it was less biased towards a specific label and had only four incorrect labels out of 16 predictions.

5.4. Partial dependence

A graphical representation called a *partial dependency plot* can be employed to depict the functional relationship between a select set of input variables and the corresponding predictions. These plots convey the manner in which predictions rely on the values of pertinent input variables. For example, one may observe a linear escalation in the probability of FOG across successive experiments using a partial dependency plot.

The analysis presented in Fig. 4 reveals that the Gaussian Process establishes a connection between the participants with FOG and the healthy subjects during the initial TUG test trial. Notably, the Decision Tree Classifier exhibits a distinct dependency pattern for the "0" label (individuals without FOG) compared to those with FOG. Moreover, the dependency on the "0" label displays a downward L-shaped trend, while the "1" label remains constant. For a more comprehensive understanding of Partial Dependence, please consult (Kabul, 2022).

5.5. Calibration curve

Comparisons between the calibrated probabilistic predictions of a binary classifier are performed using *calibration curves*, often referred to as *reliability diagrams*. These curves display the true frequency of the

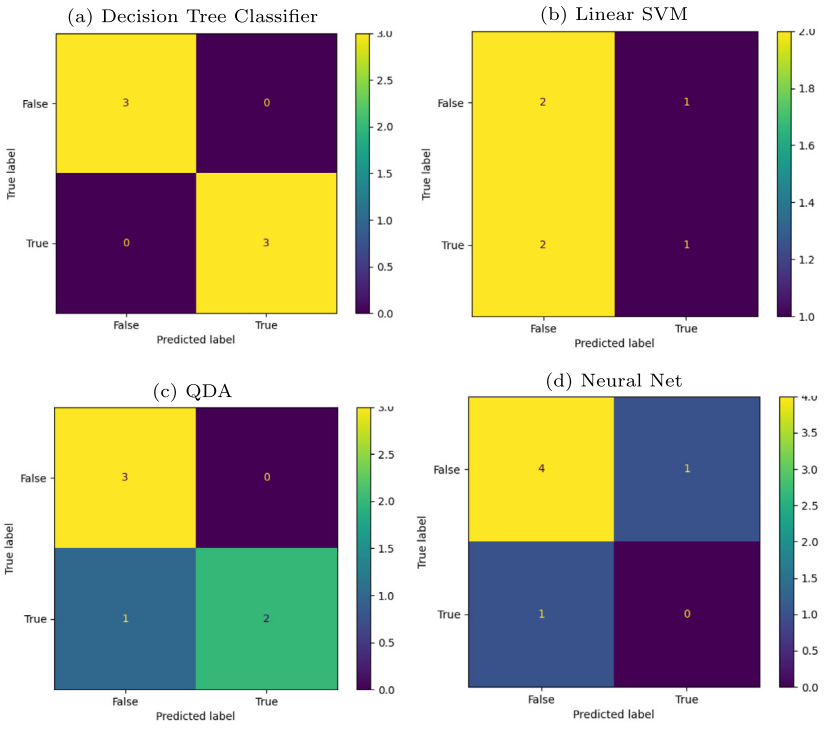


Fig. 2. Confusion matrices of single trials. These were selected due to having the highest numbers of true positives and true negatives.

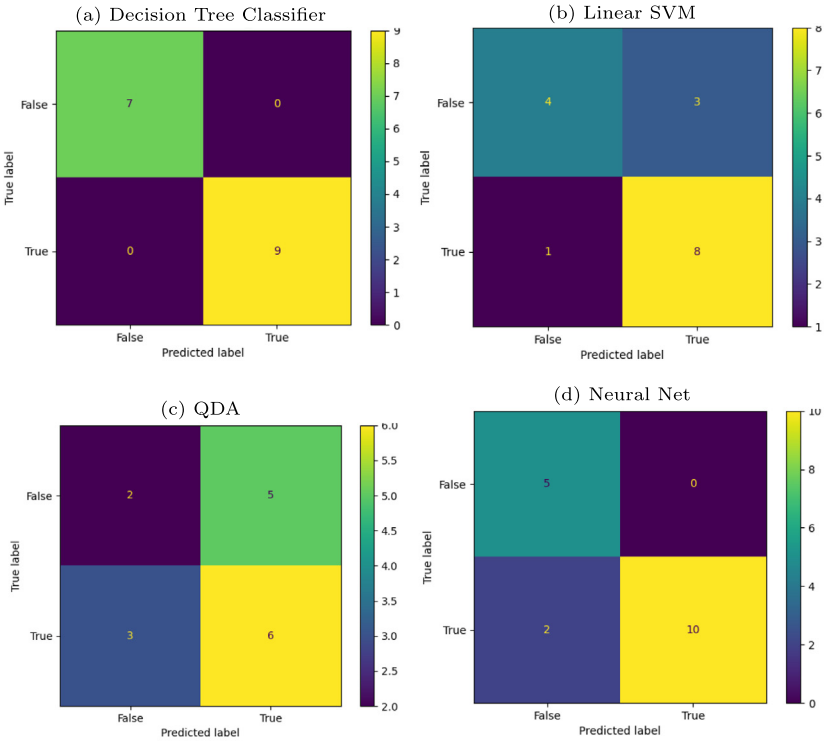
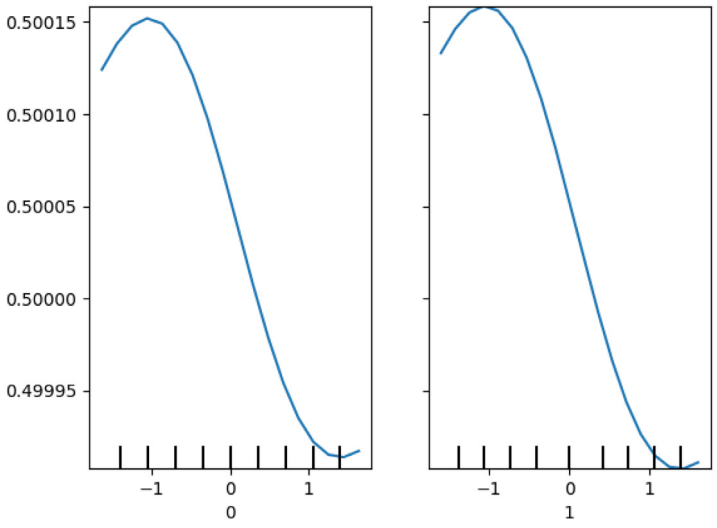


Fig. 3. Confusion matrices of all trials combined, for the same classifiers as in Fig. 2.

positive label against its predicted probability by binning the predictions, with each bin's average expected probability represented on the x-axis.

In Fig. 5, the top two figures pertain to the analysis of the Ada Boost Classifier and Decision Tree Classifier for a single trial. Within these two figures, it is evident that the line corresponding to the mean predicted probability and the fraction of positives deviates from the calibrated line (dotted-line). However, as the number of samples increases, the congruence between the two lines improves, as illustrated by the bottom two figures. In these lower figures, the Decision Tree Classifier

(a) Gaussian Process for TUG – trial 1, for participants with FOG and healthy subjects.



(b) Decision Tree Classifier for TUG – all trials combined.

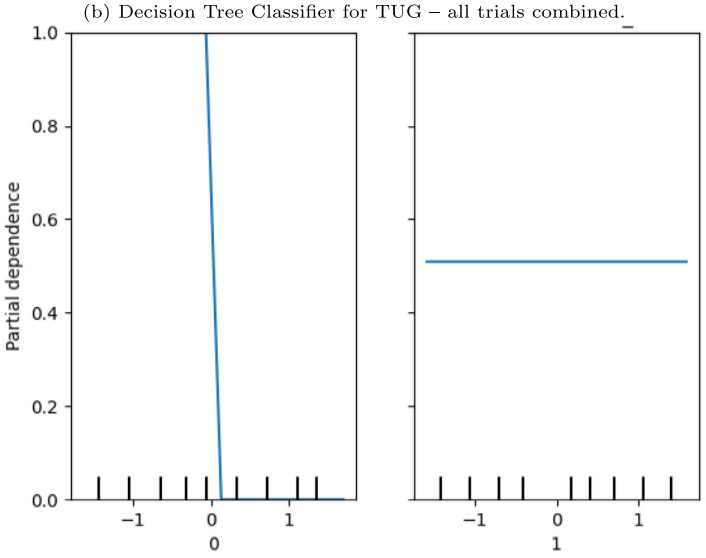


Fig. 4. Partial dependency plots for TUG.

exhibits perfect calibration in conjunction with the predicted probability of the outputs, indicating that the expected outputs (generated by its model) correspond closely to the actual classes.

6. Conclusions and future work

This study has presented a dataset containing records from various experiments and trials of both PD patients and healthy volunteers in order to illustrate the distinguishing features of PD from healthy individuals, including with respect to the ON and OFF states for medication for FOG. Motivation for the creation of such a dataset is provided in Chen (2012), in which it was stated that many unanswered questions remain about the ON-medication state of FOG.

In using this data, two main problems were considered, namely classifying PD patients from healthy individuals, and assessing the degree to which medication was effective. In considering these problems, several classifiers were implemented and their performance analyzed. Although most of the classifiers performed well with the extracted features, the multilayer perceptron and decision tree classifiers showed the most consistent results for both challenges. Notably, Decision Tree also performed well for both problems. When considering individual tasks, the best results were obtained for the Timed Up and Go (TUG) task, in which four of the classifiers obtained scores of 100% on all metrics.

Features extracted from all patients before and after taking their medication are analyzed, and their effectiveness is calculated. However, demonstrating whether a medicine is helping an individual patient during the diagnosis is not evaluated. Indeed, conducting a patient-wise analysis could provide valuable insights into the individual variability of Parkinson's disease symptoms and their response to treatment. By focusing on each patient separately, the unique gait characteristics and the effectiveness of different treatments could be evaluated and compared. This information could assist in tailoring personalized treatment plans and enhancing the overall management of the disease. A patient-wise analysis may also give an opportunity to further understand the patient's symptoms, leading to more effective treatment strategies in the future.

Furthermore, feature importance can be calculated to better understand gait impact in Parkinson's disease. This can be achieved through methods such as feature selection, dimensionality reduction, and model-based methods. The results can provide valuable insight into the most effective features for analysis and the underlying biological mechanisms of the disease.

Finally, in detecting the time periods during which FOG occurs, it would be useful to identify features that occur near that time period.

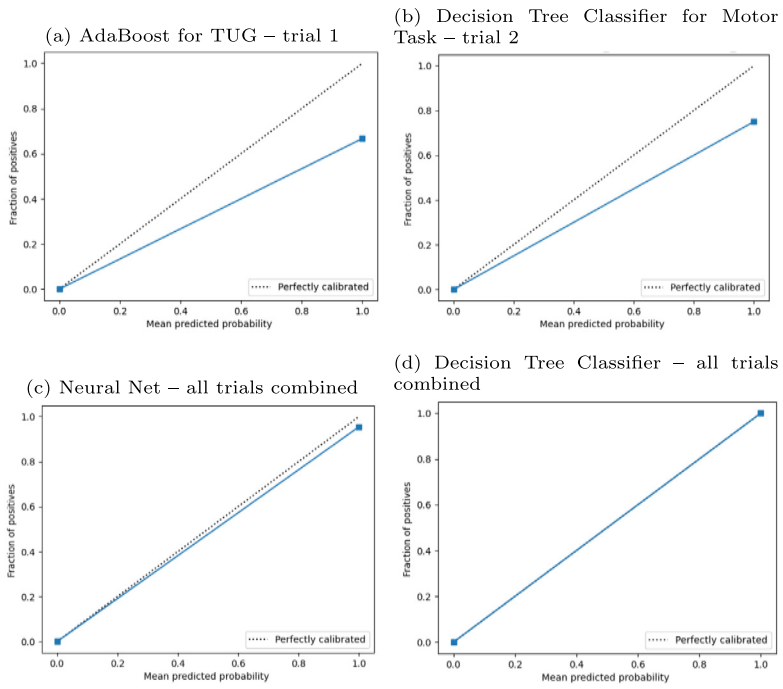


Fig. 5. Calibration curves for single trials and all trials combined.

The findings of this study can be used to design a system that alerts patients when they are about to have FOG. This type of information may assist individuals in preventing falls and injuries. Furthermore, FOG detection can be used in real-time to send cues to patients in order to assist them. Finally, the presence of FOG may indicate the need for medication adjustment; thus, the results can be incorporated into clinical practice.

Declaration of competing interest

To the best of my knowledge there is no conflict of interest associated with submitting this manuscript to your journal.

Acknowledgments

The present work was carried out with the support of the National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel, Brazil (CAPES – Program CAPES/DFATD-88887.159028/2017-00, 88887.343650/2019-00, Call no. 34/2017, Program CAPES/COFECUB-88881.370894/2019-01) and the Foundation for Research Support of the State of Minas Gerais, Brazil. A. O. Andrade is fellow of CNPq, Brazil (304818/2018-6 and 302942/2022-0). It was also supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Almeida, Q.J., Lebold, C.A., 2010. Freezing of gait in parkinson's disease: A perceptual cause for a motor impairment? J. Neurol. Neurosurg. Psychiatry 81, 513–518.
Andrade, A.O., 2017. Novas tecnologias aplicadas à saúde: Integração de areas transformadoras a sociedade. UERN: Açu, Brazil 53, 978857621164–8.
Andrade, A.O., et al., 2017. Pelvic movement variability of healthy and unilateral hip joint involvement individuals. Biomed. Signal Process. Control 32, 10–19.
Balaji, E., Brindha, D., Balakrishnan, R., 2020. Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. Appl. Soft Comput. 94, 106494.
Bartels, A.L., et al., 2003. Relationship between freezing of gait (FOG) and other features of Parkinson's: FOG is not correlated with bradykinesia. Clin. Neurosci. 10 (10), 584–588.
Beck, E.N., Ehgoetz Martens, K.A., Almeida, Q.J., 2015. Freezing of gait in Parkinson's disease: an overload problem? PLoS One 10 (12), e0144–986.
Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152.
Cabral, A.M., Andrade, A.O., 2020. Cheat sheet: Sistema de monitoramento de distúrbios do movimento.
da Capato, T.T., Domingos, J.M.M., de Almeida, L.R.S., 2015. Versãos em Português da Diretriz Europeia de Fisioterapia para a Doença de Parkinson: Desenvolvida por vinte associações profissionais europeias e adaptada para Português Europeu e do Brasil.
Chen, R., 2012. Paradoxical worsening of gait with levodopa in Parkinson disease. Neurology 78 (7), 446–447.
Dorsey, E.R., et al., 2007. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68 (5), 384–386.
Drotár, P., et al., 2014. Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease. Comput. Methods Programs Biomed. 117 (3), 405–411.
Drotár, P., et al., 2016. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif. Intell. Med. 67, 39–46.
Elkurdi, A., et al., 2018. Gait speeds classifications by supervised modulation based machine-learning using kinect camera. Med. Res. Innov. 2 (4), 1–6.
Ellis, R.J., et al., 2015. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. PLoS one 10 (10), e0141694.
Eskidere, Ö., Ertaş, F., Hanilçi, C., 2012. A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Syst. Appl. 39 (5), 5523–5528.
Goetz, C.G., et al., 2008. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov. Disorders 23 (15), 2129–2170.
Hathaliya, J.J., et al., 2022. Parkinson and essential tremor classification to identify the patient's risk based on tremor severity. Comput. Electr. Eng. 101, 107946, URL https://www.sciencedirect.com/science/article/pii/S0045790622002245.
Hausdorff, J.M., et al., 2003. Impaired regulation of stride variability in Parkinson's disease subjects with freezing of gait. Exp. Brain Res. 149, 187–194.
Hughes, J.A., Brown, J.A., Khan, A.M., 2016. Smartphone gait fingerprinting models via genetic programming. In: 2016 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 408–415.
Hughes, J.A., Houghten, S., Brown, J.A., 2019a. Descriptive symbolic models of gaits from Parkinson's disease patients. In: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, pp. 1–8.
Hughes, J.A., Houghten, S., Brown, J.A., 2019b. Models of Parkinson's disease patient gait. IEEE J. Biomed. Health Inf. 24 (11), 3103–3110.
Hughes, J.A., Houghten, S., Brown, J.A., 2020. Gait model analysis of Parkinson's disease patients under cognitive load. In: 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 1–8.
Kabul, I.K., 2022. Interpret model predictions with partial dependence and individual conditional expectation plots. https://blogs.sas.com/content/subconsciousmusings/2018/06/12/interpret-model-predictions-with-partial-dependence-and-individual-conditional-expectation-plots/, Accessed: 07-09-2022.