Tampere University

Pinja Koivisto

SYMPTOM ANALYSIS OF PARKINSON’S DISEASE UTILIZING MACHINE LEARNING METHODS

Bachelor’s thesis
Faculty of Medicine and Health Technology
Milla Juutinen
May 2023

ABSTRACT

Pinja Koivisto: Symptom analysis of Parkinson’s disease utilizing machine learning methods
Bachelor’s thesis
Tampere University
Bachelor's Program in Biotechnology and Biomedical Engineering
May 2023

While monitoring Parkinson’s disease progression or observing the everchanging severity stage of the disease, the patients are keeping symptom diaries and making regular visits to the neurologist clinic for evaluation. The diaries are based on patients own memories which tend to be unreliable in addition to the burdensome clinical appointments. Therefore, the research is focused on automatizing the burden with the help of machine learning classifiers. These classifiers are trained to either recognize the current severity stage of a patient or make a prediction about future outcomes, such as the progression rate of the disease or a freezing of gait event. The data on which the classifiers are trained with is gathered via wearable sensors that attain several gait parametrics from different walking tasks or daily activities conducted.

This thesis presents several studies conducted during the years of 2020–2023 which aim to develop a machine learning algorithm to classify the correct state of the patient according to the disease stage, or predict medical outcomes before their occurring. Their performance metrics are evaluated, especially regarding their accuracy, sensitivity and specificity results. Additionally, this thesis introduces background of gait analysis and machine learning methods. The changes in gait that Parkinson’s disease inflicts are discussed alongside the clinical criteria used in evaluating the changes and patient’s condition.

This thesis is a literature review, which aims to find the best possible machine learning algorithms for symptom analysis of Parkinson’s disease. It concludes that comprehensive conclusions are difficult to draw, since the algorithm performance can be analysed with several different metrics. Even though most of the algorithms gained adequate results, the research still includes several limitations to solve before the algorithm can be validated for clinical use as a symptom monitoring system.

Keywords: Parkinson’s disease, machine learning, motor symptoms, gait analysis, wearable sensors, automatization, symptom evaluation, performance metrics

TIIVISTELMÄ

Pinja Koivisto: Parkinsonin taudin oireiden arviointi koneoppimismenetelmien avulla
Kandidaatintutkielma
Tampereen yliopisto
Bioteknologian ja biolääketieteen tekniikan kandidaattiohjelma
Toukokuu 2023

Parkinsonin taudin etenemisen seuranta perustuu potilaiden omiin oirepäiväkirjamerkintöihin. Lisäksi taudin jatkuvasti muuttuvaa vakavuusastetta seurataan säännöllisesti neurologin klinikalla. Päiväkirjat perustuvat potilaan omiin muistikuviin, jotka ovat yleensä epäluotettavia ja klinikalla käynti raskasta. Siksi tutkimus keskittyy taakan automatisointiin koneoppimismenetelmien avulla. Nämä algoritmit koulutetaan joko tunnistamaan taudin nykyinen vakavuusaste tai ennustamaan tulevia tuloksia, kuten taudin etenemisnopeutta tai kävelykyvyn jäätymistä. Tietoja, joilla koneoppimisalgoritmeja koulutetaan, kerätään puettavien sensoreiden avulla. Nämä keräävät dataa useista eri kävelyparametreista, jotka saadaan talteen erilaisia kävelytestejä hyödyntäen.

Tässä työssä esitellään useita vuosina 2020–2023 tehtyjä tutkimuksia, joiden tarkoituksena on kehittää koneoppimisalgoritmeja, jotka luokittelevat potilaan oikeaan vakavuusastekategoriaan tai ennustavat lääketieteellisiä tuloksia ennen niiden ilmenemistä. Algoritmien suorituskykymittareita arvioidaan erityisesti tarkkuuden, herkkyyden ja spesifisyyden suhteen. Lisäksi työssä taustoitetaan kävelyanalyysin periaatteita, puettavia sensoreita sekä yleisimpiä koneoppimismenetelmiä, joita tutkimukset ovat käyttäneet. Parkinsonin taudin myötä kävelyyn kohdistuvia muutoksia käsitellään ja potilaan tilan arvioinnissa käytettyjä kliinisiä kriteerejä esitellään.

Tämä työ on kirjallisuuskatsaus, jonka tavoitteena on löytää parhaat mahdolliset koneoppimisalgoritmit Parkinsonin taudin oireiden analysointiin. Tuloksista voidaan päätellä, että kattavaa johtopäätöstä on vaikea tehdä, koska algoritmien suorituskykyä voidaan analysoida useilla eri mittareilla. Vaikka suurin osa algoritmeista saivatkin onnistuneita tuloksia, tutkimukset sisälsivät silti useita rajoituksia, jotka ovat ratkaistava ennen kuin algoritmi voidaan validoida kliiniseen käyttöön oireiden seurantajärjestelmänä.

Avainsanat: Parkinsonin tauti, koneoppiminen, motoriset oireet, kävelyanalyysi, puettavat sensorit, automatisaatio, oireiden arviointi, suorituskykymittarit

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neuronal Networks
AUC	Area Under the Curve
BC	Bayes Classifier
CNN	Convolutional Neuronal Networks
DL	Deep Learning
DOF	Degree of Freedom
DT	Decision Tree
EC	Ensemble Classifier
ELA	Ensemble Learning based Adaboost
ER	Ensemble Regressor
FOG	Freezing of Gait
GAN	Generative Adversarial Network
GB	Gradient Boosting
GMB	Gradient Boosting Machine
HC	Healthy Control
HY	Hoehn and Yahr
IMU	Inertial Measurement Unit
KNN	K-Nearest-Neighbour
LDA	Linear Discriminant Analysis
LOOCV	Leave-One-Out Cross-Validation
LR	Logistic Regression
LSTM	Long Short-Term Memory
LWRF	Locally Weighted Random Forest
MDS	Movement Disorder Society
MEMS	Micro-Electro-Mechanical System
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes
NN	Neuronal Networks
PD	Parkinson’s Disease
RBF-SVM	Support Vector Machines with Radial Basis Kernels
RF	Random Forest
SVM	Support Vector Machine
TUG	Timed-Up-and-Go
UPDRS	Unified Parkinson's Disease Rating Scale
VGRF	Vertical Ground Reaction Force
XGB	Extreme Gradient Boosting

1. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disease wherein the amount of dopamine secreting neurons in the midbrain's substantia nigra starts to decrease. The cause of the cell death of dopaminergic neurons is unknown. Since dopamine is responsible for coordinating normal movements, the loss of dopamine in the system leads to the main movement signs of Parkinson's, which include gait and balance problems, bradykinesia, rigidity and tremor. Bradykinesia refers to slowness of movement. (Ronken and Scharrenburg, 2002; Triarhou, 2013)

PD is the second most common neurogenerative disease after Alzheimer's disease. It can be considered as a multifactorial disease since the main cause of development remains unclear. Additionally, the symptoms are heterogeneous and include non-motor symptoms such as sleep disorders and psychiatric disturbances, as well. The main risk factor for PD is age, but both genetics and environmental factors can play a part in developing the disease. (Stoker et al., 2018)

As the disease progresses, the symptoms become more severe. The treatment predominantly relies on drugs that aim either to restore the dopamine levels or to act on post-synaptic dopamine receptors (Stoker et al., 2018). Levodopa is currently the most effective treatment for PD, to which nearly all the patients have a good response to. However, prolonged use of levodopa leads to motor fluctuations, ultimately effecting the quality of life (Simuni and Pahwa, 2009). Due to the progression of the symptoms, PD will slowly lead towards motor disability. There is yet no cure found for Parkinson's disease.

Nowadays, the usual way of symptom monitoring and evaluation of the changes are conducted via patient diaries and visits to the neurological clinics for visual observations. These procedures are relying on patient's own memory and notes that usually are unreliable. (AlMahadin et al., 2020) Additionally, the clinical evaluation utilizes their own standards to classify the current severity stage, which have a chance of misclassification and low efficiency since several of the criteria are descriptive symptoms. These can fail to provide a quantified diagnostic basis. (Balaji et al., 2020) With machine learning, the possibility of automatic monitoring at home could be enabled, without the need for clinical visits. Patients would be wearing sensors which gather data according on gait and motor fluctuations. An automatic system could gather the data and help the clinician via data-based decision making, rather than visual observations.

Machine learning algorithms proposed in the research are aiming to automatize the burden of both the patient and the clinician. These algorithms are trained to either classify or predict certain outcomes. For example, they can classify a patient according to the gait data collected to the proper severity scale of the disease. Additionally, the classification can be done based on the motor state: whether a patient is in ON state or OFF state medication-wise. As for prediction, the progression rate of the disease could be assessed, or detecting an episode of freezing a few seconds before it occurs.

This thesis focuses on gait assessment and monitoring of daily changes of patients already diagnosed with Parkinson's disease. In addition to introducing some gait measurement devices, different machine learning algorithm performances are evaluated for the symptom analysis. The intention of this literature review is to give insight, where does the technology of evaluating daily symptom variations and machine learning algorithm accuracy stand today when focusing on gait assessment of patients living with Parkinson's disease. Several studies published during the years of 2020–2023 are reviewed to find the best performing machine learning methods for symptom evaluation.

The second section introduces different types of sensors used to attain gait parameters for autonomized symptom evaluation. Additionally, the changes in gait of Parkinson's patients and different methods used to analyse those changes are described both from the research and clinical point of view. Section three briefly introduces the most common machine learning methods used in the section fours analysis of research studies. Section four presents all the studies found related to the symptom evaluation of Parkinson's disease according to four different subsubsections: stage detection, severity assessment, motor symptom fluctuations and symptom monitoring. Lastly, section five concludes the authors thoughts concerning the results and future evolvement.

2. GAIT MEASUREMENTS

Gait, the manner of walking, is interpret as a learned complex motor skill that facilitates movement. While gait can be conducted automatically without constant effort, it requires integration of motor control, balance, cognition and musculoskeletal function in order to function properly. The ability to walk is considered as a basic part of the quality of life, to which many disorders as well as aging has influence. (Jankovic and Tolosa, 2015, p. 622)

Walking is comprised of repetitious sequence of limb motions to move the body forward while retaining stance stability. Each sequence involves a series of interactions between two lower limbs and the whole-body weight. As the body moves forward, one limb provides support while the other limb advances to a new support site followed by reversing the roles. Both feet are in contact with the ground when transferring the body weight from one limb to the other. The simplest way of describing gait cycle is according to the variations of the foot-ground-contact. (Perry and Burnfield, 2010, p. 3)

Each of the repetitive gait cycles can be divided into two periods: stance and swing. Stance is the term used to describe the entire period while the foot is on the ground. Stance begins with initial contact in which a person initiates ground contact with their heel. Swing applies to the period where the foot is in the air. Swing begins as the foot is raised from the ground (toe contact off). (Perry and Burnfield, 2010, p. 4)

The gait cycle starts by lifting another limb forward while the rear limb is extended. The foot rolls on the ground while holding most of the body weight as a part of the stance phase. As the body mass moves forward, the toes lift from the ground, initiating the swing phase. During the swing phase, the leg moves forward after the hip is flexed while the rear knee is initially flexed and later extended for the leg to reach the ground. The foot is dorsiflexed in order to avoid the toes contacting the ground. The gait cycle ends when the heel touches the ground again. (Jankovic and Tolosa, 2015, p. 622)

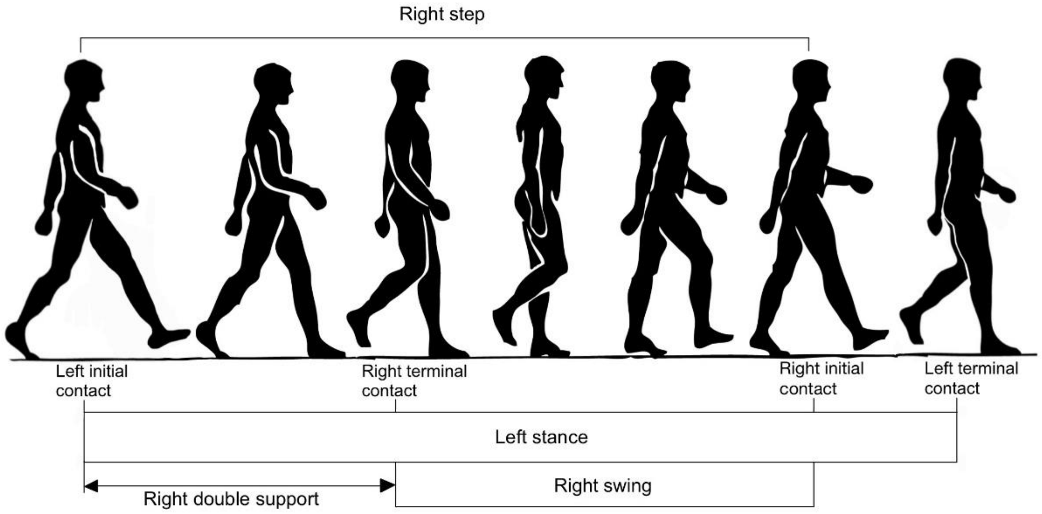


Figure 1. *The gait cycle. Modified from (Jankovic and Tolosa, 2015, p. 623).*

The proportion of each cycle is determined by speed of walking in addition to the physical state and equilibrium. Accelerated walking correlates with shortening of all the phases, but mainly in the double-limb support time. Physical weakness and aging are amongst those that increase double-limb support time. (Jankovic and Tolosa, 2015, p. 622)

2.1 Accelerometers and gyroscopes

To measure physical activity such as gait, motion sensors are used. Accelerometers are lightweight, portable and small non-invasive devices that measure the movement of the body in the matter of acceleration, which means change in the speed in respect of time. They provide insight to the intensity and volume of locomotion. Accelerometers can be either uniaxial, biaxial or triaxial according to the number of planes. (Varum and André, 2011, p. 3)

According to Kong and Bassett Jr. (2015), most of the accelerometers used in physical activity measurements are piezoelectric sensors which detect acceleration in one to three orthogonal planes: vertical, mediolateral and anteroposterior. When the piezoelectric sensor undergoes acceleration, a voltage signal is generated that is proportional to applied acceleration (Kong and Bassett Jr., 2015). Accelerometers can detect only the acceleration of the body part where they are placed. Uniaxial accelerometers are susceptible to acceleration in a single plane, usually vertical. Therefore, to truly measure one's acceleration in different directions, multiple sensors in orthogonal planes are crucial. For the optimal measurement of gait in all planes of movement, triaxial accelerometers are usually used. (Varum and André, 2011, p. 179)

Gyroscopes are sensors that are able to detect angular velocity. They can measure the turn rates caused by changes in position with respect to inertial space (Collin et al., 2019). Micro-Electro-Mechanical System (MEMS) gyroscopes are motion sensors that measure the angular motion of the target. They measure the rate of the rotation around the particular axis: 1-axis, 2-axis and 3-axis. (Passaro et al., 2017)

MEMS usually use a vibrating mechanical element to sense and detect the angular velocity. When a MEMS gyroscope is experiencing rotation, a force called Coriolis force will act and cause it to move in a direction perpendicular to its vibrating direction. This movement is proportional to the rotation speed converting it to electrical signals thus transferring energy. A microcontroller can read the signals and determine the angular velocity of the subject. (Zhuang and Zhou, 2020, p. 2)

2.2 Inertial measurement units

Accelerometers and gyroscopes are inertial sensors, often combined to form an inertial measurement unit (IMU). IMUs are small wearable devices that are widely used in gait assessment. They are capable of conducting an evaluation on large amount of steps and allow an objective evaluation of gait and movement disorders outside of the clinical environment (Washabaugh et al., 2017). Since IMU based sensors can measure data from where they are placed, for a gait assessment, usual positioning of devices is in the lower limb area as well as in the lower back.

IMUs are mainly used in devices to measure orientation, velocity and gravitational force. The earlier types of IMUs consist of accelerometer and gyroscopes with three degrees of freedom (DOF) to measure data from three axes. DOF dictates the number of independent parameters in a system. Newer types of IMUs have an additional magnetometer, usually triaxial as well, which measures the bearing magnetic direction. Magnetometers can be calibrated to the gyroscope data thus improving the reading of the gyroscope. (Ahmad et al., 2013)

2.3 Change in the gait patterns of people with Parkinson’s disease

Gait disorders appear in almost all cases of PD, often advancing towards loss of mobility and increased mortality (Ebersbach et al., 2013). Motor symptoms such as tremor, bradykinesia (slowness of movement) and stiffness are most common symptoms of Parkinson’s disease that influence the gait patterns of patients. Problems with gait becomes a burden as the disease progresses, which affects independence and quality of life (Mirelman et al., 2019). In advanced stages of PD, gait disorders become more complex, including motor blocks (freezing of gait) and festination (Ebersbach et al., 2013).

Often the first motor symptoms, reduced amplitude of arm swing and smoothness of movement are specific for patients with Parkinson’s. In early to middle stages of PD, the gait of patients becomes bradykinetic and step length decreases compared to age-matched healthy controls. In addition, irregular timing of steps and loss of rhythm become more prominent as the disease progresses from early stage. Range of motion in knees, ankles and hips begin to reduce during walking, particularly in the late-stance phase of the gait cycle. As ambulation becomes less automatic, many gait variations become noticeable and overemphasized, for example when patients are asked to conduct other activities such as searching for keys of their pockets while walking. (Ebersbach et al., 2013; Mirelman et al., 2019)

As the disease progresses to advanced stage, the changes in gait worsen: motor blocks, festination, and problems with gait initiation begins to appear. Additionally, reduced postural and balance control leads to high risk of falling. Muscle force declines, minimizing motor capacity and leading to the need for assistance devices. (Ebersbach et al., 2013; Mirelman et al., 2019)

Freezing of gait (FOG) is an episodic gait disturbance which can be defined as sudden episodes of inability to initiate or maintain movement or make a turn (Jankovic and Tolosa, 2015) and part of the late-stage symptom range of Parkinson’s. FOG episodes tend to be brief, followed by resuming of normal gait. Experiencing festination while walking can be described as the tendency to move forward rapidly but ever smaller steps, connected with the centre of gravity falling forward over the stepping feet. Festination may lead up to an episode of freezing while stepping becomes increasingly shorter and faster ending up to a complete blockage. Festination may occur as its own, but mostly in patients experiencing FOG episodes. (Ebersbach et al., 2013)

2.4 Measurement methods to extract gait parameters

The clinical point of view in assessing PD patients’ disease severity or progression is derived from observations based on clinical criteria, such as Unified Parkinson’s Disease Rating Scale or Hoehn and Yahr scale, which are more detailed in the Section 4. In a clinical assessment, a patient is accompanied by a specialist when conducting several motor tasks and/or answering questions. Each of the tasks are scored based on the patient’s ability to perform them. Conclusions can be drawn for the disease severity, progression, treatment efficiency, response and side effects. The follow-up observations are done by the patients themselves, in a form of a symptom diary or own memories. These may include marks related to the number of FOG episodes, or other major fluctuations of the motor state as well as documentation of the use of medication. (AlMahadin et al., 2020)

As for research, the patients usually have wearable sensors containing an IMU attached to different locations on the body, usually legs, to obtain gait data. In addition to IMU’s, other sensor may be used, such as vertical ground reaction force sensors which can detect the force applied to the ground, for the collection of gait data. This data is then being processed and analysed with for example, machine learning methods, to see, whether an algorithm can predict the same results as a specialist – without a need of a specialist him/herself attending.

Several different tactics can be used to gather relevant gait data for research. The simple one is conducting a walking test, in which a patient walks a predetermined distance or time. The walking task can contain also turning or multitasking. A usual test made when evaluating a FOG event is called Timed-Up-and-Go (TUG) test. The TUG test is conducted by taking time for how long it takes from a patient to rise from a chair, walk 3 meters, turn and sit again (Browne and Nair, 2019). TUG test can also be used to predict the risk of falls. For monitoring purposes, the patients could be conducting daily living activities at home, while the gait data is being collected via inertial sensors.

3. MACHINE LEARNING METHODS

Machine learning (ML) is programming computers to optimize a performance criterion using testing data or past experience. The ML model may be descriptive to gain knowledge from the inputted data or predictive to make predictions concerning the future outcome, or both. Machine learning utilizes the theory of statistics in building mathematical models, since the object is to make inference from a sample. (Alpaydin, 2014, p. 3) ML techniques can be divided into three categories: unsupervised learning, supervised learning and reinforcement learning, of which unsupervised and supervised techniques are introduced.

In supervised learning, the algorithm uses a training set and known output responses to develop a model that creates reliable predictions for the new input data. It includes training a model with these labelled input and output values. The correct values of the output are provided by a supervisor. (Alpaydin, 2014, p. 11; Himani et al., 2021, p. 43) On the contrary, unsupervised learning does not rely on labelled data nor does it have a supervisor providing correct output values. They are trained with raw and unlabelled data. Their aim is to find irregularities in the input. (Alpaydin, 2014, p. 11) Unsupervised learning can be used for clustering tasks while supervised can be used for classification and regression. Different ML methods, such as deep learning, can be both supervised and unsupervised, depending on the use case.

The following subsection 3.1 presents the most used ML classifiers according to the Section 4, which mostly belong to the supervised learning category, namely to classification. The classification method separates the datasets into various classes or categories by adding a label. It assigns each of the data points to a specific group based on a certain criterion. Classification is done in order to perform predictive analysis on the dataset. (Himani et al., 2021, p. 44)

3.1 Machine learning classifiers

Decision trees (DT) are a multistage decision processes, in which different subsets of features are used at separate levels of the tree. The model moves through the tree from the roots to the leaf according to a ‘yes’ or ‘no’ structure. The root node is considered as a node that has no incoming lines, thus being at the top of the tree. Internal nodes are ones with one incoming line, and two or multiple outgoing lines. The leaf nodes have one incoming line and no outgoing lines. Each of the non-terminal nodes represent one of the features and the lines coming from that node represent a value or possibly a set of values, which that feature could take. A class label is associated with each of the leaf nodes. (Himani et al., 2021, p. 44; Webb and Copsey, 2011, p. 323)

Random forest (RF) models are comprised of multiple de-correlated trees collected which are then averaged to reduce variance. Each of the trees are constructed using a bootstrap sample: if the database contains *n* patterns, *n* samples are taken, with replacement, to generate a bootstrap set of size *n*. As a result, approximately two-thirds of the patterns are used for training the classifiers and the remaining one-third retained for testing. The procedure is repeated to all the trees using different bootstrap sample of the data. A majority vote of the trees that did not contain the pattern in the bootstrap sample used for their construction (the one-third of trees), is obtained and the classification of the pattern is achieved by using the majority vote. The RF method is a combination of bagging and decision tree classifiers. (Hastie et al., 2009, p. 587; Webb and Copsey, 2011, pp. 389–390)

Support vector machine (SVM) method is used for constructing an optimal separating hyperplane between two separated classes, or to a high-dimensional feature space. Margin describes the sum of distances from the separating hyperplane to the closest sample of each of the classes. The maximal margin, in other words the largest distance, determines the hyperplane. The larger the margin is, the better generalization error of the classifier defined by the hyperplane. However, in many real-world problems there is no linear boundary separating the classes which would make the search of an optimal hyperplane meaningless. To these problems, non-linear SVM methods are used. This method transforms the input features nonlinearly to a space in which the linear methods can be applied. (Hastie et al., 2009, pp. 417–420; Webb and Copsey, 2011, pp. 249–250, 291)

Deep learning (DL) is a subset of machine learning but differs from ML based on in the depth of its analysis and the kind of automation it provides. DL tries to mimic the human brain functionality. It processes data by its computing units, called neurons, which are arranged into ordered sections, called layers. The foundational technique DL utilizes is the neural network, described next. (Mueller and Massaron, 2019, chap. 1)

Neuronal networks (NN), also known as the **Artificial neuronal networks (ANNs)**, are comprised of several neurons (also called units) with each neuron linking to the inputs and outputs of other neurons. A neural network can work with complex data since it allows multiple inputs to flow through multiple layers of processing to produce countless outputs. Neurons take weighted values as an input, sum them, and provide the summation as a result. Alike actual neurons of the brain, each of the paths activate only when they have a chance of answering to the question posed with inputs: after receiving weighted values, they sum them and use an activation function to evaluate the results, which transforms the result in a nonlinear way. For example, the activation function can release a zero value if the input does not achieve a certain threshold. (Mueller and Massaron, 2019, chap. 7)

The NN architectures have different layers, each one having its own weights. Additionally, each layer has different number of neurons. The number of neurons between two layers dictates the number of connections. Weights correlate to the strength of the connection between neurons in the network. (Mueller and Massaron, 2019, chap. 7)

Convolutional neuronal networks (CNNs) are alike ANNs, they are composed of neurons that self-optimize through learning. However, they take images as an input. Each neuron receives an input and performs an operation. The entire network will express a single perceptive score function: the weight. Besides image inputs, another difference between the CNN and ANN architectures is that the layers comprised of neurons are organized into three dimensions: height, width and depth. (O’Shea and Nash, 2015)

The CNN architecture consists of three layers: convolutional layer, pooling layer and fully connected layers. The input layer holds the image’s pixel values. Convolutions work by operating on small image parts, also called moving image windows, across all image channels simultaneously. The window starts from the upper left corner of the image, moving from left to right and from top to bottom. The route across the whole image is called a filter, or a kernel, and implied a complete transformation of the image. Convolution filters can detect an edge or enhance certain characteristics of an image, such as colour. The convolutional layer thus transforms the original image using filtering. The pooling layers receive outputs from convolutional layers and simplify them, thus downsizing the data flowing through the neuronal network by reducing parameters and computational complexity. The fully connected layer performs the classification based on previous layers and their filters with addition to the features extracted from them. Neurons within this layer have connections to all of the outputs of the previous layers (Mueller and Massaron, 2019, chap. 10; O’Shea and Nash, 2015)

Different types of CNN models have been used in the Section 4 analysis. These contain for example GoogleNet, ResNet and AlexNet. They differ from each other in the basis of filter amount, the number of layers in both convolutional and pooling layers or types of pooling layers used to name a few.

3.2 Cross-validation

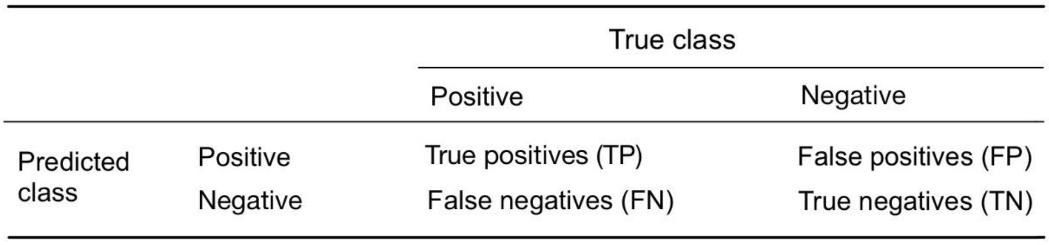
Overfitting is a term used to describe a ML algorithm that is too complex, hence it may model noise in the training set (Webb and Copsey, 2011, p. 6). This is because the classifier does not only learn the underlying function but also the noise (Alpaydin, 2014, p. 39). It will lead to good performance with the training data, but poor performance with unseen data, also known as the validation or testing data. Cross-validation can reduce the possibility of overfitting.

One cross-validation method is *K*-fold, where the dataset *X* is divided randomly into *K* parts of the equal size. To generate each pair (training and testing), one of the *K* parts is kept out as the testing set and the remaining *K* – 1 parts are combined to form the training set. Usually, the *K* is either 10 or 30, of which 10-fold cross-validation is also known as tenfold. An extreme version of *K*-fold is called leave-one-out or LOOCV. In LOOCV, the given dataset of *N* instances, only one instance is left out as the testing set and the training set uses *N* – 1 instances. This will result in *N* separate pairs by leaving out different instances at each iteration. (Alpaydin, 2014, p. 559)

3.3 Classifier performance metrics

Performance of a classifier can be assessed in many ways. For a practical application, different machine learning classifiers may be implemented to choose the best one. Many performance metrics can be calculated from a confusion matrix. (Webb and Copsey, 2011, p. 404) Table 1 presents a 2 x 2 confusion matrix for two classes, positive and negative.

Table 1. *2 x 2 confusion matrix.*



		True class	
		Positive	Negative
Predicted class	Positive	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)

True positives (TP) refer to the number of subjects of the positive class which are correctly predicted by the classifier to be in the positive class. False positives (FP) present the number of subjects of the negative class which are incorrectly predicted into the positive class. (Webb and Copsey, 2011, p. 405)

In this thesis, the main performance metrics under interest are accuracy, sensitivity, specificity and area under the receiver operating characteristic (ROC) curve, known as the AUC value. The performance metrics are presented in the Table 2.

Table 2. *Performance metrics, where P = TP + FN, N = FP + TN*

Accuracy (Acc)	$\boldsymbol{\frac{TP+TN}{P+N}}$
Sensitivity (Sens)	$\boldsymbol{\frac{TP}{P}}$
Specificity (Spec)	$\boldsymbol{\frac{TN}{N}}$

The ROC curve is a visual presentation of a classifier's performance. It is a plot of the true positive rate on the y-axis against the false positive rate on the x-axis. Area under the ROC curve, AUC, can gain values between 0–1. The closer the number is to one, the better the classifier's performance is in distinguishing between negative and positive classes. (Webb and Copsey, 2011, pp. 415–418)

4. SYMPTOM EVALUATION

The traditional measures for assessing the current state of a Parkinson's disease patient symptom-wise are conducted in a clinic by a specialized neurologist. The measures contain multiple motors tasks for evaluating the functional condition of a patient. With the help of machine learning classifiers and wearable sensors obtaining gait data from the subjects, the long process of observing changes in the gait parameters could one day be automatized.

Qualitative rating scales are used to assess the functional condition of Parkinson's disease patients. Most referred ones are the Unified Parkinson's Disease Rating Scale (UPDRS) and Hoehn and Yahr scale (HY). The UPDRS acts as a rating tool in measuring the course of the disease in patients. In 2001, the Movement Disorder Society (MDS) sponsored a critique towards the UPDRS, identifying number of ambiguities, weaknesses and areas which needed to reflect current scientific developments. Based on the critique, a new version called MDS-UPDRS was created. The full MDS-UPDRS contains several items/tasks, divided across part I to part IV. The MDS-UPDRS rates 65 items in comparison to 55 on the original UPDRS. The original UPDRS scale has 48 items with 5 possible response options (0 = normal, 1 = slight, 2 = mild, 3 = moderate, and 4 = severe) and 7 with yes/no responses to describe a patient's state. The third part (MDS-UPDRS III) assesses motor tasks, such as gait, postural stability, arising from chair and freezing of gait. (Goetz et al., 2008)

The Hoehn and Yahr scale defines broad categories of motor function. The scale contains 1-5 stages. Progression in these stages has been found to correlate with declining of motor ability and quality of life as well as neuroimaging studies of dopaminergic loss. A patient scoring 1 in the HY scale has shown only unilateral involvement, usually with mild or no functional disability. On stage 5, the patient is in bedrest or wheelchair unless aided. Comparing the two mentioned scales for PD symptom severity assessment, the MDS-UPDRS scores for all 4 parts increase with every HY stage. However, the HY does not provide information concerning non-motor symptoms of PD like the MDS-UPDRS score does in part I. (Bhidayasiri and Tarsy, 2012; Skorvanek et al., 2017)

The following subsections review different machine learning classifier abilities to detect correctly the UPDRS or HY rating of a patient. Additionally, the differences in a patients' motor state and its fluctuations as well as detection of freezing of gait is assessed. Lastly, the performance of ML classifiers are evaluated for the purpose of long-time monitoring of the progression in PD state and patients' gait. The tables referred in Sections 4.1 – 4.4 excluding the subsection 4.3.1 are included in the appendix.

4.1 Stage detection

As a progressive neurogenerative disease, Parkinson's has multiple stages corresponding to the HY scale of 1–5. The research papers found for this section had mostly PD patients in mild to moderate stages of 1–3, to conduct the study about an automatic stage detection. Since in stage 5 the symptoms are so severe, that the patient is in bedrest or needs a wheelchair, no walking tasks can be performed. Thus, this stage is not represented in the research. For stage detection analysis, three research papers were found with the most common machine learning method being random forest. The important characteristics of each of the studies has been collected into the Table 6.

All studies conducted a walking test or TUG test with varying sensors or data acquisition systems to assess gait parameters. Mirelman et al. (2021) used body-fixed sensors containing tri-axial accelerometers and gyroscopes placed on the subject's lower back and ankles while Ferreira et al. (2022) used cameras to calculate various spatial-temporal gait parameters. Varrecchia et al. (2021) had an optoelectronic motion analysis system with the patients having 22 markers covered with aluminium powder over prominent bony landmarks to detect movement. These studies conducted a walking test. The method Seedat and Aharonson (2020) used is an instrumented walker with build-in accelerometer, force sensors and encoders, which calculate distance to capture kinematic data of the subject's gait. While collecting gait data, the walker simultaneously supports walking. To capture the data, this study conducted a TUG test on patients.

Regarding stage detection, it is important to have diagnosed PD patients across different HY stages to attend in the study for the algorithm to be reliable. As stated before, mostly patients in HY scale 1–3 were attending with an exception of stage 4 in the study of Seedat and Aharonson (2020) with total of 67 PD attending and Varrecchia et al (2021), 76 PD patients. Ferreira et al. (2022) had a similar amount of PD patients, 63 in total. The largest group of 332 patients were on the study conducted by Mirelman et al. (2021). In addition to the PD patients, healthy controls (HC) attended as well in all the studies. While Mirelman et al. (2021); Seedat and Aharonson (2020) focused solely on PD stage detection, Ferreira et al. (2022) and Varrecchia et al. (2021) conducted a two-step study, first step being the discrimination of healthy controls from PD patients, before moving into the stage detection.

After the data has been collected, the chain of study continues with data processing, feature selection, classification and lastly evaluation. As for feature selection, Ferreira et al. (2022) used 44 features in the first step which was classifying PD from HC. The feature amount selected to the second step, is assumed to be 12 according to the first step's classifiers results. From these, the two most relevant features found in rating the severity were: stride width variability and step double support time. Total of 134 gait features was used in Mirelman et al.'s (2021) study alongside subject demographics. For pairwise HY stage assessment, for example HYI vs. HYII, 18 features were selected. Seedat and Aharonson (2020) selected 211 features for the classification. Varrecchia et al. (2021) obtained 18 different kinematic parameters. Most of the classification models in these three studies include a RF classifier. As stated by Seedat and Aharonson (2020), random forest suits well in studies with small datasets with addition to allowing feature importance to be quantified in respect to their discrimination power.

The evaluation of the classification accuracy is based on its correct patient categorizing according to the HY score. The classification accuracy of the random forest classifier in the study conducted by Seedat and Aharonson (2020) varied in between of 90–96 % depending on the feature set to be classified. Ferreira et al. (2022) determined the AUC value to two different classifiers: random forest and naïve Bayes. Of these two, random forest scored higher AUC value of 0.786 than naïve Bayes 0.771. The classification model of Mirelman et al. (2021) differs from the others. They used random under-sampling boosting (RUSboost) classification model which uses decision trees and quadric discriminant analysis resulting in two classifiers per classification task. Random forest permutation importance was used in feature selection. With the selected features, the RUSboost classifier's AUC ranged between 76–90 % with mean sensitivity and specificity values ranging between 72–83 % and 69–80 % respectively. The artificial neuronal networks used by Varrecchia et al. (2021) gained a mean performance rate of 66.16–77.2 % for correct stage classification. Additionally, mean sensitivity and specificity values range from 66 % to 77 % and from 85 % to 91 %, respectively.

The only studies that addressed the limitations were Mirelman et al. (2021) and Varrecchia et al. (2021). In Mirelman et al.'s study, the PD participants gait recordings were measured during their ON-medication state, with no data from the OFF-state. Especially in the more advanced stages of PD, the marked motor response fluctuation may differ during OFF-state in medication. Future work is required to better establish the findings in a longitudinal study. As for Varrecchia et al. (2021), limitations include lack of comparison between the results with gold standard clinical measures for gait assessment. Additionally, there is room for improvement in the model performance.

4.2 Severity assessment

Alike in stage detection, severity assessment can be done via referring Parkinson's disease state to the HY scale. In addition, the following section contains references to the UPDRS rating. Even though the scaling approach is similar to the stage assessment, this section covers studies that focus on severity assessment with different method to gather the gait data.

All the following studies had their gait data from an open source, called Physionet. The data contained signals gathered from vertical ground reaction force (VGRF) sensors placed in the soles of each foot. The VGRF sensors measure force in newtons, that is applied to the ground. In total, 16 sensors, eight in each foot measured gait parameters when subjects conducted a walking test. Two different datasets are reported in these studies, both from Physionet. The first contains 279 gait recordings from 93 Parkinson's disease patients and from 73 healthy controls. The second has 306 gait recordings from 93 PD patients and 72 healthy controls. Additionally, the database contains labels to which HY scale and/or UPDRS rating the patient belongs to. All the important characteristics of each of the studies have been gathered to the Table 7.

Although the gait data used is merely the same in the studies, different features and classification methods have been selected. The most common machine learning classifiers were DT and SVM. Others include DL algorithms such as AlexNet, k-nearest neigbour, ANNs such as convolutional neuronal networks. The amount of gait features selected to the classification varied from nine to 34.

For the most common ML methods, Balaji et al. (2020) reported accuracies for four different classifiers based on statistical and kinematic features selected. Decision tree classifier gained the best results, both cumulative statistical and kinematic features 99.4 % accuracies followed by support vector machine 97.6 % for statistical and 99.4 % for kinematic feature accuracies. These two classifiers also had the same AUC value of 0.99. Other classifiers used were Bayesian classifier (BC) and ensemble classifier (EC), whose classification results are in Table 7.

Khera and Kumar (2022) used a hybrid model that first used DT classifiers that predict gait, followed up with ensemble regressors (ER). They are a combination of several decision trees and used to determine the severity of PD. The parameters are then tuned using grid search and cross-validated with tenfold and LOOCV methods. The LOOCV classifier gained 99.39 % accuracy while tenfold got as high as 99.9 % in accuracy.

Cantürk (2021) used SVM and KNN classifiers. Additionally, so called lasso and relief were used to reduce the number of features. Lasso stands for least absolute shrinkage and selection operator, which assigns feature coefficients and modifies them while relief ranks features according to their importance (Cantürk, 2021). The study used binary classification to discriminate PD patients and multiclass to predict the disease severity based on gender. The best average accuracies gained were all with lasso. For multiclass, SVM lasso gained the accuracy of 98 % while for binary classification KNN lasso had the best accuracy of 99 % for all subjects combined. Cantürk (2021) also tested the classifiers gender-wise using the same classifiers and obtained as great results than with the whole dataset: 100 % female, 99 % male with SVM. The last study to interpret DT and SVM classifiers was Wang et al. (2022). In addition, they also used KNN, naïve Bayes and ensemble learning based adaboost (ELA) classifiers. The best accuracy of 96.69 % was gained with SVM classifier.

The following studies interpret deep learning and neuronal networks for the severity classification. Aşuroğlu and Oğul (2022) used a hybrid deep learning model that consists of convolutional neuronal networks and locally weighted random forest (LWRF) architectures. The hybrid classifier gained the accuracy of 99.5 % with 34 time and frequency domain features. Veeraragavan et al. (2020) used ANN classifier with 34 gait features including spatiotemporal and kinematic features. The accuracy of the classifier was 87.1 %.

Lastly, Setiawan et al. (2021) used four different deep learning algorithms: AlexNet, ResNet-50, ResNet-101 and GoogLeNet with 40 selected features. They separated the subjects into two sections: two-class containing HC and all PD patients and multiclass corresponding the different HY levels from zero (HC) to 3. Additionally, they divided the Physionet data according to three different datasets contributed by four different studies made. The results were also presented merely according to the sub datasets in multiple different tables, which made the overall interpretation of the results challenging. However, the best average accuracy reported in results was 96.52% using ResNet-50. Additionally, the best scores from the Setiawan et al. (2021) study's table containing accuracy, sensitivity, specificity and AUC results concerning the Physionet database as a whole, are presented in the Table 7.

Since all the studies used merely the same database, the limitations mentioned align. The amount of PD subjects in the database is relatively small with addition to the uneven distribution regarding to the UPDRS/HY scale. For example, the HY scale 4 and 5 are not represented at all and between 1-3, there is an uneven number of patients in the database. However, this could have been a conscious choice to preserve the safety of patients in a poor health condition. One solution could be to increase the sample size of PD patients with an even distribution according to the severity rate. Other solution is to use the Synthetic Minority Oversampling Technique (SMOTE) to increase the sample size. Additional limitations mentioned are lack of other symptom types, such as non-motor symptoms, which could be considered to improve the prediction rate.

4.3 Motor symptom fluctuations and freezing of gait

Parkinson's disease patients are often prescribed with levodopa to ease the motor symptoms. With medication, patients experience cycles in the severity of their motor symptoms, labelled as ON state for the time the drug is active and OFF state when the effect wears off and the symptoms worsen (Ramesh and Bilal, 2022). Dyskinetic state (DYS) refers to the period when involuntary movements are noticeable. Dyskinesias occur as a complication of the levodopa treatment. Since these motor fluctuations are affecting the patients quality of life, ideally, the PD patients would remain constantly in the ON state, resembling better motor function without experiencing the OFF or dyskinetic states (Pfister et al., 2020).

Freezing of gait is one of the most troublesome gait effecting symptoms of PD. It occurs involuntary as an episodic absence of forward progression of the feet. FOG episodes as well as motor state fluctuations are currently monitored by patients' themselves in a form of diaries and questionnaires. To evoke a FOG episode during gait measurements, the patients must be conducting tasks which include turning or stopping from command.

4.3.1 Motor state detection

Different motor states described earlier include ON, OFF and DYS states. When the patients are visiting a neurologist in a clinic, they often are either in ON or OFF state, or transitioning between the states. The neurologist can determine the state using the UPDRS exam. Progression-wise, usually the HY scale is assessed. Without a proper, continuous monitoring of the motor states, the dynamic between them is impossible to capture. (Ramesh and Bilal, 2022)

For the motor state assessment, two studies have been found. Pfisher et al. (2020) conducted a study containing 30 PD patients doing daily activity tasks, collecting 11,567 minutes of accelerometer data. The six accelerometers used were attached to the subject's both wrists and feet as well as on the back and chest. Ramesh and Bilal (2022) had conducted two different studies, containing 58 PD patients in total with IMU's attached around the limbs and torso while conducting a walking test. In the first study the subjects were recorded twice, for both ON and OFF states. The second study recorded the subjects up to five times over six-hour period, which is the approximate duration of a full ON/OFF cycle. Some of the patients did not experience the full cycle. Both Phisher et al. (2020) and Ramesh and Bilal (2022) recorded video from the patients for several specialist to review and determine the appropriate motor state. Details and performance metrics of all the studies are presented in the Table 3.

Table 3. *The study characteristics for motor state estimation.*

REFERENCE	THE METHOD TO COLLECT GAIT DATA	THE NUMBER OF PARTICIPANTS	CLASSIFICATION METHOD	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
(PFISTER ET AL., 2020)	11 567 minutes of IMU data from patients doing daily-life activities, a smartwatch in wrist of more affected side.	30 PD	CNN (seven layers)	76.8 (OFF), 66.7 (ON), 77 (DYS) Three-class accuracy 65.4	64 (OFF), 67 (ON), 64 (DYS)	89 (OFF), 67 (ON), 89 (DYS)
(RAMESH AND BILAL, 2022)	IMU placed on the lower back, walking test.	Study 1: 35 PD 2 visits: ON and OFF. Study 2: 23 PD up to five times over 6h period of a full ON/OFF cycle.	GAN (three layers), CNN (two layers)	100 (best CNN study 1), 78 (best CNN study 2), 100 (best GAN study 1), 100 (best GAN study 2)	-	-

For the classifier methods, both studies used neuronal networks. Phisher et al. (2020) used CNNs while Ramesh and Bilal (2022) used generative adversarial networks (GANs) in addition to CNNs. The GAN architecture consists of two neural networks: a generator and a discriminator. The discriminator network is alike a traditional CNNs: trained to output whether a sample is fake or real to minimize loss function. The generator is trained to fool the discriminator by creating fake samples. Additionally, the generator maximizes the discriminator’s loss. (Ramesh and Bilal, 2022)

The CNN model Pfisher et al. (2020) used gained highest accuracy score of 77 % for detecting DYS state, followed by OFF 76.8 % and ON 66.7 % state accuracies. The overall, three-class accuracy of 65.4 % was obtained. The data contained total of 26.8 % OFF state, 41.4 & ON state and 31.8 % dyskinetic state. Each CNN was trained with 15 randomly selected PD patients. Ramesh and Bilal (2022) did not measure the dyskinetic state. Their CNN accuracy gained 100 % for study 1 state detection and 78 % for study 2. The proposed GAN architecture outperformed the CNNs, gaining 100 % accuracies in both studies. Their models were tested with 10 study 1 development set subjects and nine study 2 subjects. The training of the classifiers was conducted on 25 subjects from study 1.

Limitations in both studies have been noticed. One includes the environment, in which the studies have been conducted. The constrained clinical environment and data collection protocol does not resemble free living activities of the patient in daily life. Thus, the current models may not generalize as well to gait measurements collected outside of the clinic. Secondly, due to the small cohort of patients, the models may not be representative of larger variety of Parkinson's disease patients and their different severity scales.

4.3.2 Freezing of gait prediction

For an automatic FOG episode detection, three studies were found. Borzi et al. (2021) had 11 PD patients wearing two IMUs attached on shins performing a TUG test. The patients were both ON and OFF state medication wise. Kleanthous et al. (2020) had 10 PD patients participating, from which eight ended up experiencing FOG episodes during the measurement. The two who did not experience FOG were excluded. The accelerometer data was gained from an ankle, thigh and trunk of a patient performing walking test which include turning and stopping. Two of the participants were ON state. Borzi et al (2023) had their data from three different datasets: REMPARK, 6MWT and ADL. The REMPARK dataset includes 21 PD patients both ON and OFF states doing different walking tasks in home environment. An IMU measuring acceleration data was attached to the left side of the waist of a patient. 6MWT dataset includes 38 PD patients with ON state and 21 HC. The participants conducted a walking test that required turning wearing triaxial accelerometer and gyroscope on the lower back. The ADL datasets had 59 PD patients who were all ON state. The participants were asked to perform walking tasks, which included turning and sitting down and up. The number of FOG episodes recorded in each of the studies can be found on Table 8. All the studies included video raters of experts to confirm the FOG episodes.

In each of the studies different machine learning classifiers were used. Borzi et al. (2021) utilized decision trees for feature selection and support vector machines for FOG classification with 10 time domain features and six frequency domain features extracted. Additionally, leave-one-subject-out method was used for validation. In pre-FOG detection, they utilized k-nearest neighbour, linear discriminant analysis and logistic regression with 10-fold cross-validation. Kleanthous et al. (2020) utilized random forest, extreme gradient boosting, gradient boosting, support vector machines using radial basis functions and neural network classifiers with 30 final selected features. They focused on the prediction of FOG and considered three following time periods: 2, 3 and 4 s, prior to the onset of FOG. Borzi et al. (2023) had convolutional neuronal network architecture as a classifier for FOG detection.

For Borzi et al. (2021) the accuracy for FOG detection varied between 92–96.3 %, depending on the validation method and whether the patent was in ON or OFF state. Generally, higher results were obtained with 10-fold cross-validation for patients in OFF state. After training the algorithm with PD patients in ON state and testing on PD patients on OFF state, and vice versa, the accuracy range dropped slightly, scoring 89 % for ON state and 92.6 % for OFF state. As for pre-FOG recognition, the accuracy range was between 44.9–94.7 %. The best results were provided by SVM and linear discriminant analysis (LDA) classifiers. When combining the two classifiers, the accuracy gained 91.7 % and 92.9 % for ON and OFF states respectively. The best accuracy scores for pre-FOG with LDA and SVM classifiers separately are presented in the Table 8 with 2s window length.

Kleanthous et al. (2020) used multiple different classifiers with detection accuracy ranging between 77–97 % for transitioning. The highest accuracy results for transitioning, FOG and walk events combined are presented in the Table 8. The predictor values (pred.) in parentheses in Table 8 describe number of top features used in addition to the period (s) before a FOG event. Additionally, the sensitivity and specificity values of summative performance results can be obtained from the Table 8. The best performance was obtained with support vector machines with radial basis kernels (RBF-SVM), achieving sensitivity values of 72.34 %, 91.49 %, 75.00 %, and specificity values of 87.36 %, 88.51 % and 93.62 % regarding FOG, transition and normal activity classes, respectively.

Lastly, the CNN model of Borzi et al. (2023) was able to correctly identify 91.2 % of the FOG episodes with an average of 68.7 % of FOG detected in each episode. The highest AUC score of 0.955 was obtained for the training set, followed by 0.947 and 0.946 for validation and test sets respectively. The training set included 12 of the PD patients while validation and test sets included four and five PD patients respectively. These all sets are from the REMPARK datasets, since it had the greatest number of recorded FOG episodes. The results obtained by testing the algorithm on the 6MWT dataset are presented in Table 8.

What comes to limitations, the size of Kleanthous et al. (2020) and Borzi et al. (2021) datasets are rather small. Additionally, Kleanhous et al. (2020) conducted their study in a controlled environment which may not resemble daily living activities. As for Borzi et al. (2023), there were no FOG episodes recorded in the ADL dataset. Additionally, the main dataset, REMPARK, included most of the FOG episodes but the activity label was not available.

4.4 Symptom monitoring

As of today, the assessment of gait fluctuations and FOG episodes of PD patients are still based on questionnaires or scales, such as the UPDRS. Additionally, long-time monitoring is based on the patients' own diaries which may not be a reliable source. In clinical visits, scores of scales are calculated, but they are based on at the time which is not sufficient to accurately monitor the symptoms (Li et al., 2022). Therefore, the research for new methods for long-time monitoring of the PD patients' motor state and disease progression as well as analysing the illness is needed. This subsection introduces five studies which three of them assess gait monitoring, one demonstrates potential for estimating PD severity scores from home and final study to develop a predictive model for an individual's PD progression rate. Details and performance metrics of all the studies are presented in the Table 9.

For gait monitoring systems, Li et al. (2022) had 14 PD and 8 HC attending to the gait measurements, which included walking tasks and TUG test. Participants were wearing an IMU located on the lateral side of the ankles and force-sensitive insoles. Five PD patients and five healthy controls were attending Ilesan et al. (2022) study, wearing pressure sensors in the insoles and two EMG channels clustered into a foot biomechanics assessment module to track the lower-limb muscular activation pattern. Additionally, an accelerometer was placed to the subject's wrist. Steady-state walking tasks were acquired for gait data collection. Popescu et al. (2022) proposed a wearable device in a form of a bracelet, which has EMG sensor and accelerometer in it. The bracelet can be connected to a cloud. Unknown number of PD patients recruited conducted several walking tasks and usual activities for four days. Additionally, they conducted a TUG test and dual-task test for the PD patients. On the last day, the PD participants were on OFF state. Their study proposes an eHealth monitoring system with a deep learning model that can predict the patient's response to levodopa.

Hssayeni et al. (2021) developed an algorithm for estimating the UPDRS III scale. They had 24 PD patients attending, wearing two sensors that measure acceleration and gyroscope data on the most affected wrist and ankle. The measurement started while the subjects were in their OFF state. Fifteen of the subjects conducted daily living activities in four rounds spanned for four hours. After the first round, the subjects resumed their daily medication intake. The other nine patients cycled through multiple stations in home-like environment, also doing daily living activities while first being OFF state and later in ON state. For the nine subjects, two hours of continuous recording was gathered.

Raval et al. (2020) had gathered gait data from 160 PD patients. Six movement sensors containing triaxial accelerometer, magnetometer and gyroscope were located in the subjects on each ankle and wrist, the lower back and the upper chest. They conducted an extended TUG test (iTUG) and iSway, which gave measures such as jerk and sway area. The participants' individual UPDRS III rate and its change was followed for 24 months.

Several different machine learning classifiers were used. Li et al. (2022) compared random forest, logistic regression and gradient boosting for the classification performance with six temporal domain and eight spectral domain features extracted. Ilesan et al. (2022) utilized convolutional neuronal networks with several architectures: MobileNet, EfficientNetB0 and Xception. Another study to utilize neural networks was Hssayeni et al. (2021) with a dual-channel long short-term memory (LSTM) for hand-crafted features, 1D Convolutional Neural Network (CNN-LSTM) for raw signals, and 2D CNN-LSTM for time–frequency data. In addition, Raval et al. (2020) used XGBoost and feed forward neural networks models. Lastly, Popescu et al. (2022) implemented the AlexNet deep learning model for severity estimation of the motor state.

All three of Li et al. (2022) machine learning classifiers gained great results, which are summarized in Table 9. The best accuracy result of 97.31 % was obtained with gradient boosting classifier. Ilesan et al. (2022) gained the best results with MobileNet model with 95 % accuracy, 90 % sensitivity and 96 % specificity. Hssayeni et al. (2021) reported the results with correlation and MAE values when using an ensemble of the three deep learning models. A high correlation of $\rho = 0.79$ ($p < 0.001$) and low MAE = 5.95 was obtained. Singularly tested, the best correlation value $\rho = 0.70$ was obtained with 1D CNN-LSTM for raw signals and the best MAE 6.85 for dual-channel LSTM for hand-crafted features. The NN model of Raval et al. (2020) outperformed the XGBoost models in every case. The best model performance was thus obtained with the NN model, using clinical measures to predict the 2-year percent change in the UPDRS III score. The NN model explained 37% of the variance in the target, with a PPV of 71% in identifying fast progressors. Three of the postural and gait stability feature sets explained 10% or more of the variance in the 2-year MDS-UPDRS part III score. The performance metrics for severity estimation of motor state in Popescu et al. (2022) study with AlexNet obtained 84 % accuracy for bradykinesia estimation and 90 % for tremor. The proposed architecture gained a high accuracy of 96.5 % in gait variability when analysing three cohorts: adults, elderly and PD patients through wearable insoles. Additionally, accuracy of 91 % was obtained for analysing the gait symptoms in different PD patient severity stages.

Limitation-wise, most of the studies other than Raval et al.(2020), had a small dataset of subjects participating in the study. This may affect the reliability of the different classification models. When it comes to Ilesan et al. (2022), in their study the participant's complained about the worn-in discomfort addition to the discomfort of the wires interfering with their mobility. Their study was also conducted in a small room which likely inhibited the walking as well. Popescu et al. (2022) did not specify the number of each subjects in their cohorts: elderly, adults and PD patients. The only study to conduct a long-term measurement of the subject's state of disease for over two years was Raval et al. (2020).

5. CONCLUSIONS

Parkinson's disease is a non-curable, progressive disease with complicated symptoms. The disease state detection and progression observations merely lie on clinical visits and patients own written notes on complications and medical intake. This thesis brought up several studies, which aimed to develop machine learning algorithms to automatically detect the state of the disease, monitoring the symptoms and predicting outcomes such as the freezing of gait -episode. Several different algorithms were used by the selected research papers but all of them did not report the results with the same performance metrics, which makes it difficult to draw a comprehensive conclusion concerning their suitability in symptom evaluation.

However, five machine learning classifiers that got the highest performance metrics in terms of accuracy, sensitivity and specificity are presented in the Table 4. Additionally, these five were used in more than one study. The area under the curve values were lacking in many reports, thus it is better to leave the metric out of the final conclusions. In terms of accuracy, most of the reported algorithms gained above 90 % results. Only a couple of the classifiers scored 60–70 %, which could be explained by the lack of most relevant features fed to the classifier, for example. Two classifiers gained a 100 % accuracy: the best generative adversarial networks and convolutional neuronal networks algorithms utilized in motor state detection, both from the same study.

Table 4. *Performance metrics of the five best ML methods.*

METHOD	USED IN	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
SVM	Severity assessment FOG prediction	91.3 – 99.4	68.4 – 99.6	97.8 – 99.8
KNN	Severity assessment FOG prediction	94 – 98.5	87.95	95.98
DT	Severity assessment FOG prediction	93.98 – 99.4	87.95 – 99.6	95.98 – 99.8
CNN	Motor state detection FOG prediction Symptom monitoring	65.5 – 100	64 – 88.5	67 – 89
RF	Stage detection FOG prediction Symptom monitoring	89.4 – 96	-	-

It should be noted that all the studies did not report every performance metrics listed. This is seen in KNN's case, where two studies reported the accuracy results but only one of them presented the sensitivity and specificity results. Additionally, random forest was a popular method, but none of the studies reported the sensitivity and specificity values. As for CNN's, one study got only 65.5–77 % results in terms of accuracy, which drops the success range lower.

Since convolutional neuronal networks gained merely great results, also the different architectures of traditional CNN's that gained above 90 % accuracy are presented in the Table 5. In addition, the high scoring GAN method is listed. These methods were used only once, but they could be studied more in terms of symptom analysis because of their high-performance scores. In ResNet's case there were no specification on which of the two (ResNet-50 or ResNet-101) the best average scores were achieved.

Table 5. *Performance metrics of modified CNN architectures that scored high accuracy.*

METHOD	USED IN	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
CNN + LWRF	Severity assessment	99.5	98.7	99.1
RESNET	Severity assessment	94.58 (multi-class) 96.63 (two-class)	92.08 94.46	95.60 97.69
MOBILENET	Symptom monitoring	95	90	96
GAN	Motor state detection (ON/OFF)	100	-	-

To conclude what was stated before, many of the classifiers can succeed in their performance, achieving over 90 % accuracy, some even close to 100 %. However, a single study achieving 100 % accuracy with their classifier does not guarantee ongoing success in further studies, nor does it validate the classifiers position amongst the best. Therefore, different machine learning classifiers and their performance in analysing the symptoms of Parkinson's disease needs to be studied more.

Regarding future studies, if the aim is to develop an automatic symptom monitoring system, the research should be moved from clinical data collection protocol to the home environment where movement is freer, thus being more erratic and diverse. An algorithm developed with data taken in clinical environment may not generalize well outside of the clinic, thus leading to poor performance. Additionally, the amount of data and variety of subjects taking part in the study must be larger to avoid bias or misinterpretation. Most of the studies foresee a promising path towards the ML method usage in monitoring systems. For this to be possible, the measurements should be long-term and continuous to gather data both from ON and OFF state as well as transition for a system to be reliably assessing the symptoms, which change over the course of the medical cycle and progression of the disease.

REFERENCES

Ahmad, N., Ghazilla, R.A.R., Khairi, N.M., Kasi, V., 2013. Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications. Int. J. Signal Process. Syst. 256–262. https://doi.org/10.12720/ijsps.1.2.256-262

AlMahadin, G., Lotfi, A., Zysk, E., Siena, F.L., Carthy, M.M., Breedon, P., 2020. Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. BMC Neurol. 20, 419. https://doi.org/10.1186/s12883-020-01996-7

Alpaydin, E., 2014. Introduction to machine learning, Third edition. ed, Adaptive computation and machine learning series. MIT Press, Cambridge, Massachusetts.

Aşuroğlu, T., Oğul, H., 2022. A deep learning approach for parkinson's disease severity assessment. Health Technol. 12, 943–953. https://doi.org/10.1007/s12553-022-00698-z

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. https://doi.org/10.1016/j.asoc.2020.106494

Bhidayasiri, R., Tarsy, D., 2012. Parkinson's Disease: Hoehn and Yahr Scale, in: Bhidayasiri, R., Tarsy, D. (Eds.), Movement Disorders: A Video Atlas: A Video Atlas, Current Clinical Neurology. Humana Press, Totowa, NJ, pp. 4–5. https://doi.org/10.1007/978-1-60327-426-5_2

Borzì, L., Mazzetta, I., Zampogna, A., Suppa, A., Olmo, G., Irrera, F., 2021. Prediction of Freezing of Gait in Parkinson's Disease Using Wearables and Machine Learning. Sensors 21, 614. https://doi.org/10.3390/s21020614

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. Artif. Intell. Med. 135, 102459. https://doi.org/10.1016/j.artmed.2022.102459

Browne, W., Nair, B. (Kichu) R., 2019. The Timed Up and Go test. Med. J. Aust. 210, 13-14.e1. https://doi.org/10.5694/mja2.12045

Cantürk, İ., 2021. A computerized method to assess Parkinson's disease severity from gait variability based on gender. Biomed. Signal Process. Control 66, 102497. https://doi.org/10.1016/j.bspc.2021.102497

Collin, J., Davidson, P., Kirkko-Jaakkola, M., Leppakoski, H., 2019. Inertial Sensors and Their Applications. Springer.

Ebersbach, G., Moreau, C., Gandor, F., Defebvre, L., Devos, D., 2013. Clinical syndromes: Parkinsonian gait. Mov. Disord. 28, 1552–1559. https://doi.org/10.1002/mds.25675

Ferreira, M.I.A.S.N., Barbieri, F.A., Moreno, V.C., Penedo, T., Tavares, J.M.R.S., 2022. Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters. Gait Posture 98, 49–55. https://doi.org/10.1016/j.gaitpost.2022.08.014

Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A.E., Lees, A., Leurgans, S., LeWitt, P.A., Nyenhuis, D., Olanow, C.W., Rascol, O., Schrag, A., Teresi, J.A., van Hilten, J.J., LaPelle, N., 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment. Mov. Disord. 23, 2129–2170. https://doi.org/10.1002/mds.22340

Hastie, Trevor., Tibshirani, Robert., Friedman, Jerome., 2009. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition, 2nd ed. 2009. ed, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Himani, B., Balusamy, B., Poongodi, T., Khan KP, F., 2021. Machine learning and analytics in healthcare systems : principles and applications, First edition. ed, Green engineering and technology: concepts and applications. CRC Press, Taylor & Francis Group, Boca Raton.

Hssayeni, M.D., Jimenez-Shahed, J., Burack, M.A., Ghoraani, B., 2021. Ensemble deep model for continuous estimation of Unified Parkinson’s Disease Rating Scale III. Biomed. Eng. OnLine 20, 32. https://doi.org/10.1186/s12938-021-00872-w

Ileșan, R.R., Cordoș, C.-G., Mihăilă, L.-I., Fleșar, R., Popescu, A.-S., Perju-Dumbravă, L., Faragó, P., 2022. Proof of Concept in Artificial-Intelligence-Based Wearable Gait Monitoring for Parkinson’s Disease Management Optimization. Biosensors 12, 189. https://doi.org/10.3390/bios12040189

Jankovic, J., Tolosa, E., 2015. Parkinson’s disease & movement disorders. Wolters Kluwer, Philadelphia.

Khera, P., Kumar, N., 2022. Novel machine learning-based hybrid strategy for severity assessment of Parkinson’s disorders. Med. Biol. Eng. Comput. 60, 811–828. https://doi.org/10.1007/s11517-022-02518-y

Kleanthous, N., Hussain, A.J., Khan, W., Liatsis, P., 2020. A new machine learning based approach to predict Freezing of Gait. Pattern Recognit. Lett. 140, 119–126. https://doi.org/10.1016/j.patrec.2020.09.011

Kong, C.Y., Bassett Jr., D.R., 2015. The Technology of Accelerometry-Based Activity Monitors: Current and Future. Off. J. Am. Coll. Sports Med. 37.

Li, Y., Bai, Q., Yang, X., Zhou, X., Sun, Y., Yao, Z., 2022. An abnormal gait monitoring system for patients with Parkinson’s disease based on wearable devices, in: 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Presented at the 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–6. https://doi.org/10.1109/CISP-BMEI56279.2022.9980005

Mirelman, A., Ben Or Frank, M., Melamed, M., Granovsky, L., Nieuwboer, A., Rochester, L., Del Din, S., Avanzino, L., Pelosin, E., Bloem, B.R., Della Croce, U., Cereatti, A., Bonato, P., Camicioli, R., Ellis, T., Hamilton, J.L., Hass, C.J., Almeida, Q.J., Inbal, M., Thaler, A., Shirvan, J., Cedarbaum, J.M., Giladi, N., Hausdorff, J.M., 2021. Detecting Sensitive Mobility Features for Parkinson’s Disease Stages Via Machine Learning. Mov. Disord. 36, 2144–2155. https://doi.org/10.1002/mds.28631

Mirelman, A., Bonato, P., Camicioli, R., Ellis, T.D., Giladi, N., Hamilton, J.L., Hass, C.J., Hausdorff, J.M., Pelosin, E., Almeida, Q.J., 2019. Gait impairments in Parkinson’s disease. Lancet Neurol. 18, 697–708. https://doi.org/10.1016/S1474-4422(19)30044-4

Mueller, J.Paul., Massaron, L., 2019. Deep learning, 1st edition. ed, For dummies. For Dummies, Hoboken, New Jersey.

O’Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks.

Passaro, V.M.N., Cuccovillo, A., Vaiani, L., De Carlo, M., Campanella, C.E., 2017. Gyroscope Technology and Applications: A Review in the Industrial Perspective. Sensors 17, 2284. https://doi.org/10.3390/s17102284

Perry, J., Burnfield, J.M., 2010. Gait analysis - Normal and Pathological Function, 2nd Edition. ed. SLACK Incorporation, USA.

Pfister, F.M.J., Um, T.T., Pichler, D.C., Goschenhofer, J., Abedinpour, K., Lang, M., Endo, S., Ceballos-Baumann, A.O., Hirche, S., Bischl, B., Kulić, D., Fietzek, U.M., 2020. High-Resolution Motor State Detection in Parkinson's Disease Using Convolutional Neural Networks. Sci. Rep. 10, 5860. https://doi.org/10.1038/s41598-020-61789-3

PhysioNet Databases [WWW Document], 2023. Open databases. Available: https://physionet.org/about/database/ (accessed 22.4.23).

Popescu, N., Channa, A., Ifrim, R., 2022. Neuro-cognitive Evaluations using Deep Learning and Wearable Sensorial Devices. Presented at the International Conference on New Approaches (ICNAE'22), Turkiye.

Ramesh, V., Bilal, E., 2022. Detecting motor symptom fluctuations in Parkinson's disease with generative adversarial networks. NPJ Digit. Med. 5, 138. https://doi.org/10.1038/s41746-022-00674-x

Raval, V., Nguyen, K.P., Gerald, A., Dewey, R.B., Montillo, A., 2020. Prediction of Individual Progression Rate in Parkinson's Disease Using Clinical Measures and Biomechanical Measures of Gait and Postural Stability, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1319–1323. https://doi.org/10.1109/ICASSP40776.2020.9054666

Ronken, E., Scharrenburg, G.J.M. van., 2002. Parkinson's disease. Solvay Pharmaceuticals Conferences, v. 1.

Seedat, N., Aharonson, V., 2020. Machine learning discrimination of Parkinson's Disease stages from walker-mounted sensors data.

Setiawan, F., Link to external site, this link will open in a new window, Che-Wei, L., 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, 5207. https://doi.org/10.3390/s21155207

Simuni, Tanya., Pahwa, Rajesh., 2009. Parkinson's disease, Oxford American neurology library. Oxford University Press, Oxford ;

Skorvanek, M., Martinez-Martin, P., Kovacs, N., Rodriguez-Violante, M., Corvol, J.-C., Taba, P., Seppi, K., Levin, O., Schrag, A., Foltynie, T., Alvarez-Sanchez, M., Arakaki, T., Aschermann, Z., Aviles-Olmos, I., Benchetrit, E., Benoit, C., Bergareche-Yarza, A., Cervantes-Arriaga, A., Chade, A., Cormier, F., Datieva, V., Gallagher, D.A., Garretto, N., Gdovinova, Z., Gershanik, O., Grofik, M., Han, V., Huang, J., Kadastik-Eerme, L., Kurtis, M.M., Mangone, G., Martinez-Castrillo, J.C., Mendoza-Rodriguez, A., Minar, M., Moore, H.P., Muldmaa, M., Mueller, C., Pinter, B., Poewe, W., Rallmann, K., Reiter, E., Rodriguez-Blazquez, C., Singer, C., Tilley, B.C., Valkovic, P., Goetz, C.G., Stebbins, G.T., 2017. Differences in MDS-UPDRS Scores Based on Hoehn and Yahr Stage and Disease Duration. Mov. Disord. Clin. Pract. 4, 536–544. https://doi.org/10.1002/mdc3.12476

Stoker, T.B., Stoker, T.B., Greenland, J.C., 2018. Parkinson's disease : pathogenesis and clinical aspects, Parkinson's disease. Exon Publications, Australia.

Triarhou, L.C., 2013. Dopamine and Parkinson's Disease, Madame Curie Bioscience Database [Internet]. Landes Bioscience.

Varrecchia, T., Castiglia, S.F., Ranavolo, A., Conte, C., Tatarelli, A., Coppola, G., Lorenzo, C.D., Draicchio, F., Pierelli, F., Serrao, M., 2021. An artificial neural network approach to detect presence and severity of Parkinson’s disease via gait parameters. PLoS One 16, e0244396. https://doi.org/10.1371/journal.pone.0244396

Varum, H., André, S. de B., 2011. Accelerometers: Principles, Structure and Applications. Nova Science Publishers, New York.

Veeraragavan, S., Gopalai, A.A., Gouwanda, D., Ahmad, S.A., 2020. Parkinson’s Disease Diagnosis and Severity Assessment Using Ground Reaction Forces and Neural Networks. Front. Physiol. 11, 587057. https://doi.org/10.3389/fphys.2020.587057

Wang, Q., Zeng, W., Dai, X., 2022. Gait classification for early detection and severity rating of Parkinson’s disease based on hybrid signal processing and machine learning methods. Cogn. Neurodyn. https://doi.org/10.1007/s11571-022-09925-9

Washabaugh, E.P., Kalyanaraman, T., Adamczyk, P.G., Claflin, E.S., Krishnan, C., 2017. Validity and repeatability of inertial measurement units for measuring gait parameters. Gait Posture 55, 87–93. https://doi.org/10.1016/j.gaitpost.2017.04.013

Webb, A.R. (Andrew R.), Copsey, K.D., 2011. Statistical pattern recognition, 3rd ed. ed. Wiley, Hoboken.

Zhuang, X., Zhou, L., 2020. Gyroscopes - Principles and Applications. IntechOpen, London, United Kingdom.

APPENDIX A: TABLES FROM THE SECTIONS 4.1- 4.4

Table 6. *The study characteristics for stage detection.*

REFERENCE	THE METHOD TO COLLECT GAIT DATA	NUMBER OF PARTICIPANTS	NUMBER OF FEATURES	CLASSIFICATION METHOD	AUC	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
(SEEDAT AND AHARONSON, 2020)	Instrumented walker with accelerometer, sensors and encoders. TUG test.	67 PD patients, 19 age-matched healthy controls (HC)	211	RF	-	90–96	-	-
(MIRELMAN ET AL., 2021)	Body-fixed sensors inc. tri-axial accelerometer and gyroscope. Walking test.	332 PD, 100 age-matched HC	134 and demographics, 18 pairwise	RUSboost	0.76–0.90	-	72–83	69–80
(FERREIRA ET AL., 2022)	Vicon Motion Systems® cameras. Walking test.	63 PD, 63 matched HC	44 (first step), 12 (second step)	RF, NB	0.786 (RF) 0.771 (NB)	-	-	-

(VARRECCHIA ET AL., 2021)	Optoelectronic motion system with 6 infrared cameras. 22 markers with aluminium powder. Walking test.	76 PD, 67 HC	18 kinematic	six ANNs	-	66.16–77.2	66–77	85–91

Table 7. *The study characteristics for severity assessment.*

REFERENCE	THE METHOD TO COLLECT GAIT DATA	NUMBER OF PARTICIPANTS	NUMBER OF FEATURES	CLASSIFICATION METHOD	AUC	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	OTHER METRICS
(BALAJI ET AL., 2020)	VGRF data, 16 sensors. Walking test. Source: Physionet	93 PD, 73 HC	9 (temporal + spatial)	DT, SVM, EC, BC	0.99, 0.99, 0.97, 0.88 (DT, SVM, EC, BC) for mild events	Statistical: 99.4 (DT), 97.6 (SVM), 95.1 (EC), 69.7 (BC) Kinematic: 99.4, 99.4, 95.2, 69.9 respectfully.	Statistical: 99.6 (DT), 95.2 (SVM), 91.5 (EC), 46.0 (BC). Kinematic: 99.6, 99.6, 91.6, 42.4 respectfully.	Statistical: 99.8 (DT), 99.2 (SVM), 98.2 (EC), 89.2 (BC). Kinematic: 99.8, 99.8, 98.2, 89.1 respectfully.	F-score (%) 99.25 (DT), 97.3 (SVM), 94.6 (EC), 70.7 (BC), (cumulative, statictical). F-score (%) 99.25, 99.25, 93.1, 72.1 (cumulative, kinematic)
(AŞUROĞLU AND OĞUL, 2022)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 73 HC	16 + 7 (frequency + time domain)	Hybrid deep learning model: CNN + LWRF	-	99.5	98.7	9.1	Correlation Coefficient (CC): 0.897, Mean Absolute Error (MAE): 3.009 and Root Mean Square Error (RMSE): 4.556
(KHERA AND KUMAR, 2022)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 72 HC	16 gait features	DT (first stage), ER (severity)	-	-	-	-	For LOOCV: RMSE = 0.989, MAE = 0.3921, and R^2 = 97%. For tenfold: mean RMSE of 0.977 ± 0.06, MAE = 0.3476 ± 0.024, and R^2 = 98.7%.

(CANTÜRK, 2021)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 73 HC	4096 non, 29 lasso, 959 relief for both SVM, KNN	AlexNet for feature extraction. SVM and KNN	-	99 (binary; KNN lasso), 98 (multiclass; SVM lasso)	99 (KNN lasso), 94 (SVM lasso)	98 (both KNN and SVM lasso)	F-score: 0.99, 0.95 respectively
(VEERARAGAVAN ET AL., 2020)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 73 HC	34 spatio-temporal + kinematic	ANN	-	87.1	-	-	-
(WANG ET AL., 2022)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 73 HC	24	SVM, KNN, NB, DT, ELA	-	96.69 (SVM), 93.98 (KNN), 93.67 (NB), 93.98 (DT), 95.48 (ELA)	93.37 (SVM), 87.95 (KNN), 87.35 (NB), 87.95 (DT), 90.96 (ELA)	97.79, 95.98, 95.78, 95.98, 96.99 respectively	F1 score (%): 0.934, 0.880, 0.874, 0.880, 0.910 respectively
(SETIAWAN ET AL., 2021)	VGRF data, 16 sen-sors. Walking test. Source: Physionet	93 PD, 73 HC	40	AlexNet, ResNet-50, ResNet-101, and GoogLeNet	0.9949 (ResNet-101, two-class: whole dataset), 0.9612 (Alexnet, class 3: whole dataset)	96.63 (ResNet-101, two-class: whole dataset), 97.74 (GoogLeNet, class 3, whole dataset)	95.28 (ResNet-101, two-class, whole dataset), 94.95 (AlexNet, class 3: whole dataset)	97.69 (ResNet-101, two-class: whole dataset), 98.61 (AlexNet, class 3: whole dataset)	-

Table 8. *The study characteristics for FOG prediction.*

REFERENCE	THE METHOD TO COLLECT GAIT DATA	THE NUMBER OF PARTICIPANTS	THE NUMBER OF FOG EPISODES	CLASSIFICATION METHOD	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	AUC
(BORZÌ ET AL., 2021)	Two IMUs, patients performed a TUG test.	11 PD	41 (ON state), 54 (OFF state)	**1.** Pre-FOG: KNN, LDA, LR. **2.** FOG-dect.: DT, SVM	**1.** 91.3/92.1 (SVM: ON/OFF), 91.7/94.7 (LDA: ON/OFF) **2.** 95.5 ON, 96.3 OFF (10-fold cv)	**1.** 68.4 (SVM), 66.2 (LDA) **2.** 95.9 ON, 97.1 OFF (10-fold-cv)	**1.** - **2.** 95.4 (ON), 93.5 (OFF), (10-fold cv)	-
(KLEANTHOUS ET AL., 2020)	Three acceleration data collecting sensors. The subject conducted walking, turning and doing typical daily activities.	10 PD (only 8 had FOG episodes)	237	RF, Extreme Gradient Boosting (XGB), Gradient Boosting Machine (GMB), RBF-SVM, Multilayer Perceptrons (MLP)	RF: 89.36 (30 pred., 4s) XGB: 79.1 (5 pred., 3s) GMB: 79.55 (30 and 15 pred., 4s) RBF-SVM: 79.85 (5 pred. 2s) MLP: 78.79 (both 30 and 5 pred., 4s).	87. 23 (FOG, 15 pred., 4s with GBM), 91.49 (transition, 5 pred. 3s, with RBF-SVM), 75.0 (walk, 5 pred., 3s with RBF-SVM).	87.36 (FOG, 5 pred., 3s, with RBF-SVM), 91.76 (transition, 15 pred., 4s with GBM), 96.81 (walk, 15 pred., 4s with GBM).	-

(BORZÌ ET AL., 2023)	Three different sources for gait data, collected via IMUs. Tasks contain walking, turning and sitting down/up.	REMPARK: 21 PD, 6MWT: 38 PD and 21 HC, ADL: 59 PD.	REMPARK: 1058, 6MWT: 52, ALD: 0.	CNNs (6 layers)	92.9 (6MWT).	88.4 (train), 87.9 (validation), 87.7 (test).	88.5 (train), 88 (validation), 88.3 (test).	0.955 (train), 0.947 (validation), 0.946 (test).

Table 9. *The study characteristics for symptom monitoring.*

REFERENCE	THE METHOD TO COLLECT GAIT DATA	THE NUMBER OF PARTICIPANTS	CLASSIFICATION METHOD	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	OTHER
(LI ET AL., 2022)	TUG and walking test with subject's wearing an IMU located on the lateral side of the ankles and force-sensitive insoles.	14 PD, 8 HC	RF, Logistic Regression (LR), Gradient Boosting (GB)	95.40 (RF), 89.25 (LR), 97.31 (GB)	-	-	F1-score: 0.9526 (RF), 0.8959 (LR), 0.9755 (GB) AUC: 0.9631 (RF), 0.9212 (LR), 0.9733 (GB)
(ILEȘAN ET AL., 2022)	Subjects conducted a walking task, wearing six pressure sensors in the insoles, two EMG channels and an accelerometer placed on wrist.	5 PD, 5 HC	CNN models: MobilNet, EfficientNetB0, Xception	95 (MobileNet), 85 (EfficientNetB0, Xeption)	90 (MobileNet, Xeption), 85 (EfficientNetB0)	96 (MobileNet), 80 (EfficientNetB0), 73 (Xeption)	-
(POPESCU ET AL., 2022)	Subjects conducted TUG test, walking test and dual tasking while walking. They wore 16 pressure sensors in the insoles which contain a 6-axis IMU.	three cohorts which counts 29 subjects	AlexNet DL	84 (bradykinesia), 90 (tremor), 96.5 (gait variability), 91 (gait symptoms in different severity stages)	-	-	-
(HSSAYENI ET AL., 2021)	Two wearable sensors: 3-axial accelerometer and gyroscope mounted on the most affected wrist and ankle. Fifteen of the subjects conducted daily living activities in four rounds spanned for four hours. The other nine patients cycled through multiple stations in home-like environment, also doing daily living activities.	24 PD	LSTM, 1D CNN-LSTM, 2D CNN-LSTM	-	-	-	Correlation: $\rho = 0.79$ ($p < 0.0001$) MAE: 5.95 (combined). $\rho = 0.70$ (1D CNN-LSTM), MAE: 6.85 LSTM, singularly.

(RAVAL ET AL., 2020)	Six movement sensors containing 3-axis accelerometer, gyroscope and magnetometer mounted on each ankle and wrist, the lower back, and the upper chest. The instrumented Timed-up-and-go (iTUG) and the instrumented Sway (iSway) test were conducted.	160 PD	NN, XGBoost	-	-	-	The model explained 37% of the variance in the target, with a PPV of 71% in identifying fast progressors.