Detection of Freezing of Gait in patients with Parkinson's Disease

Irene Heijink

UNIVERSITY OF TWENTE.

Medisch Spectrum Twente
een santeon ziekenhuis

A thesis submitted for the degree of Master of Science

Detection of Freezing of Gait in patients with Parkinson's Disease using a deep learning approach

Irene Bernadet Heijink

Enschede - October 26, 2022

Graduation internship performed at:
Neurocenter MST
University of Twente

Technical Medicine
Medical Sensing and Stimulation

Graduation committee:

Prof. dr. R.J.A. van Wezel Professor Biomedical Signals and Systems, University of Twente	**Chairman**
Dr. J.P.P. van Vugt Neurologist, Medisch Spectrum Twente	**Medical supervisor**
Dr. M.C. Tjepkema-Cloostermans Technical Physician, Medisch Spectrum Twente & University of Twente	**Technical supervisor**
E.C. Klaver MSc Technical Physician & PhD candidate, Medisch Spectrum Twente	**Daily supervisor**
Drs. R.M. Krol Lecturer professional behaviour, University of Twente	**Supervisor professional behaviour**
L. Rutten MSc PhD candidate Multi-Modality Medical Imaging, University of Twente	**External member**

Abstract

Introduction: Freezing of gait (FOG) is one of the debilitating symptoms experienced by patients with Parkinson's Disease, and the most common cause of falls in these patients. External cueing can help to overcome FOG. In order to enhance the user experience of cueing devices and to diminish intrusiveness and habituation to cues, on-demand cueing is desired. In addition, objective FOG detection enables monitoring and objective assessment of therapy. Therefore, automatic detection of FOG needs to be developed. In this research, the performance on the detection of FOG of three deep learning classifiers based on inertial measurement unit (IMU) data is studied.

Methods: Data from four studies with walking tasks ranging from straight walking, turning and obstacle course was combined. The dataset contains over 300 000 windows of which 8% was labelled as FOG. All experiments were recorded for video annotation of FOG. IMU measurements were performed using hardware and software of Xsens 3D motion tracking technology. The data was tested on three classification models: a CNN, MiniRocket and InceptionTime. Five fold cross-validation was applied to estimate an unbiased model performance. The models were evaluated using a receiver operating characteristic (ROC) curve. The model with the highest area under the ROC curve (AUC-ROC) was selected and a sensor evaluation was performed on this model. Sensors that were evaluated include the accelerometers of the upper legs, lower legs and feet. The best model in combination with the best sensor selection was trained on the training + validation set and tested on the hold out test set.

Results: In total, 71 unique participants were included in this study. The highest AUC-ROC was reached for the CNN trained on the acceleration data of the lower legs and feet with an AUC-ROC of 0.72, sensitivity of 73.7% (72.5 - 75.0%) and specificity of 60.8% (60.3 - 61.3%) on the test set. The mean AUC-ROC of MiniRocket was 0.10 smaller than the AUC-ROC of the CNN, and the mean AUC-ROC of InceptionTime was 0.04 smaller than the CNN. The difference in mean AUC-ROC for the sensor combinations was 0.01.

Conclusion: The classification algorithm has potential to be implemented in on-demand cueing devices and home monitoring applications for objective FOG detection. Further research is needed to optimize the model and improve the performance.

Keywords: Parkinson's Disease, Freezing of Gait (FOG), inertial measurement unit (IMU), automatic detection, time series classification, deep learning, convolutional neural network (CNN), MiniRocket, InceptionTime.

List of abbreviations

AAS	Adjusted Auditory Stroop task
Adam	Adaptive Moment estimation
AdamW	Adaptive Moment estimation with decoupled weight decay
AI	Artificial Intelligence
AUC-ROC	Area Under the Curve of the Receiver Operating Characteristic
CNN	Convolutional Neural Network
DAT-scan	Dopamine Transporter scan
FAB	Frontal Assessment Battery
FOG	Freezing of Gait
FOGQ	Freezing of Gait Questionnaire
HIVE-COTE	Hierarchical Vote Collective of Transformation-based Ensembles
IMU	Inertial Measurement Unit
IQR	Interquartile Range
MDS-UPDRS	Movement Disorders Society Unified Parkinson Disease Rating Scale
MEC-U	Medical Research Ethics Committees United
MMSE	Mini-Mental State Examination
MRI	Magnetic Resonance Imaging
N-FOGQ	New Freezing of Gait Questionnaire
ON	'On' dopaminergic state
OFF	'Off' dopaminergic state
PD	Parkinson's Disease
PPV	Proportion of Positive Values
ReLU	Rectified Linear activation Unit
ROC	Receiver Operating Characteristic
SPECT	Single Photon Emission Computed Tomography
WMO	Dutch Medical Research with Human Subjects Law

Contents

1 Introduction

Parkinson's Disease (PD) is a complex, progressive, neurodegenerative disease [1]. PD is the fastest growing neurological disorder in the world, driven by the aging population. The prevalence of over six million patients with PD worldwide in 2015, is expected to double in the next two decades [2].

Freezing of gait
PD involves disabling motor- and non-motor symptoms. One of the gait impairments experienced by patients with PD is freezing of gait (FOG). FOG is one of the most common cause of falls in PD patients [3]. FOG is defined as a "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk". Assessing and adjusting FOG treatments is challenging for medical professionals, since FOG may not manifest during clinic visits [4].

Cueing is one of the most common strategies to reduce FOG in patients with PD and is defined as temporal or spatial stimuli, which facilitate repetitive movement including gait. External cueing uses the strategy to focus attention on walking in order to overcome FOG and prevent falls. Effective cueing devices include visual stimuli such as a Parkinson walker projecting a laserline onto the ground and rhythmic auditory stimuli. However, the visibility of these devices for bystanders is a big disadvantage for patients. A more discrete device may, in addition, raise situational awareness of the patient and therefore be more safe. Tactile cueing by vibrating socks may provide the need for a subtle, safe and effective cueing device [5].

After promising results of a case study with the vibrating socks, a clinical study into the effectiveness of the socks was performed [5]. At the moment, a home monitoring study is started to evaluate the usability of the vibrating socks [6]. A prototype of the vibrating socks is shown in figure 1. In order to enhance the user experience, it would be favourable if cueing of the vibrating socks only activates at the moment FOG occurs or shortly before. In addition, Velik et al. found that the duration of FOG reduces more with on-demand cueing than with continuous cueing. On-demand cueing is thought to diminish intrusiveness and habituation to cues [7]. Furthermore, objective FOG detection enables monitoring patients at home and objective assessment of therapy. To provide the socks with feedback information about the presence of FOG, automatic detection of FOG needs to be developed.

Figure 1: Prototype of the vibrating socks, currently used in the Vibrating socks at home study. Photo from L.F.J. Nannings [8].

FOG detection
The gold standard to detect FOG is visual examination of video data [9]. However, clinicians often disagree on classifying an episode as freezing or not. A detection algorithm would be more objective and convenient [10]. Automatic FOG detection can be done based on wearable sensor data such as inertial measurement units (IMU) containing accelerometers and gyroscopes. Machine learning and deep learning algorithms to detect FOG are widely explored [11–16]. According to a review of Pardoel et al., the best performing classifiers for FOG detection are AdaBoosted decision trees, random forests, support vector machines and convolutional neural networks (CNN). The CNN shows the highest overall sensitivity, is extensively studied, and has the advantage of deep learning over machine learning that feature extraction is not needed. However, the published CNNs are trained and tested on small datasets up to 21 patients and 237 FOG episodes [11]. Considering the future application of the algorithm and patient comfort, ideally the number of sensors needed is limited.

In this study, we will implement FOG detection on a large dataset and test three different detection algorithms to improve detection performance. The first algorithm is a CNN, based on previous research on FOG detection of Abdallah et al. [14] and Ten Broeke [17]. The second algorithm is MiniRocket, a fast data mining classifier and the state-of-the-art model in time series classification [18]. The third algorithm is InceptionTime, an ensemble of deep CNN models and the best performing neural network for time series classification [19]. The choice for MiniRocket and InceptionTime is based on a review of Ruiz et al. comparing recent algorithms on 26 multivariate time series datasets [20]. All of these models will be trained patient-independently to ensure easy application in the vibrating socks.

1.1 Research aim

The aim of this research is to develop a detection model that automatically classifies FOG in PD patients. The research question is defined as follows:

How can freezing of gait in Parkinson's Disease be detected using convolutional neural networks based on IMU data?

Several steps will be made in order to develop a FOG detection model. First of all, four datasets with IMU data of PD patients will be combined to maximize the input data. Three time series classification models will be implemented, optimized and tested, namely a self-developed CNN, MiniRocket, and InceptionTime. Cross-validation is applied to accurately validate and test the performance of the models. More information on the models and types of cross-validation can be found in chapter 2. The type of IMU sensors (accelerometers and/or gyroscope) and the sensor locations of the IMU sensors will be evaluated.

1.2 Hypothesis

A time series classification model dedicated for the analysis of IMU data, will have sufficient accuracy in FOG classification. It is expected that adequate labelling of IMU data with one of the three models can be used for objective FOG detection. Several studies have shown that FOG classification can be done using CNNs [11] and part of the dataset in this study is also used as input to CNNs before [17, 21]. On the other hand, the model input suffers from the limitations of video annotation [10] and the heterogeneity of FOG [4], which complicates FOG detection. The CNN is expected to show at least equal performance to Opdams CNN with an AUC-ROC of 0.83 [21]. Regarding the complex architecture of InceptionTime and limited data set, it is expected that the more simple CNN outperforms InceptionTime. However, MiniRocket is expected to show the best performance, being the state-of-the-art time series classification method in literature [20]. Based on the pilot study of Ten Broeke, acceleration data of the knees and ankles is expected to provide valuable data input. Sensors

solely at the ankles will perform slightly worse than at the knees and ankles [17]. However, this should be re-evaluated on the full dataset.

2 Background

The first part of this chapter provides background information on PD and particularly FOG, while the second part introduces classification algorithms and machine learning techniques for detection of FOG.

2.1 Clinical background

The clinical part starts with a background in the anatomy and pathofysiology relevant for PD. Next, the diagnosis and treatment of PD is described, followed by information on the motor symptom freezing of gait.

Anatomy
The central nervous system consists of the cerebrum, diencephalon, brainstem, cerebellum, and the spinal cord. The cerebrum is divided in two hemispheres. It contains the cerebral cortex and several subcortical structures, including the basal ganglia, see figure 2 [22]. The basal ganglia are involved in movement control. It includes the striatum, pallidum, subthalamic nucleus and the substantia nigra. Several circuits traverse the basal ganglia, namely a motor loop (concerned with learned movements), cognitive loop (concerned with motor intentions), limbic loop (concerned with emotional aspects of movements), and an oculomotor loop (concerned with voluntary saccades) [23].

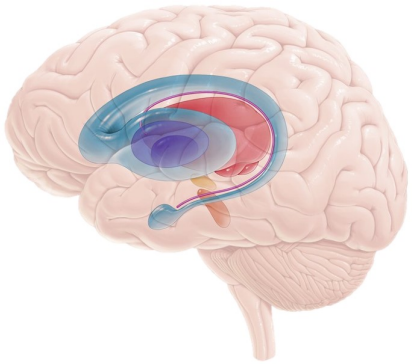


Figure 2: The basal ganglia (blue) and thalamus (red), lateral view. Picture from Andrusca et al. [24].

Pathofysiology
The primary underlying pathology of PD is degeneration of dopaminergic neurons in the substantia nigra, located in the basal ganglia [3]. The loss of dopaminergic neurons results in a loss of dopamine in the striatum of the basal ganglia [23]. Dopamine is an excitatory neurotransmitter which plays an important role in the striatal locomotor pathway. The effect of dopamine on this pathway largely explains the motor symptoms in PD [25]. In addition, dopamine is also involved in reward and reinforcement mechanisms, learning and memory mechanisms, and various other functions like attention, sleep, impulse control, and decision making [26]. The first symptoms of PD appear delayed, after about 60% of the neurons has been lost. The delay is caused by an increased dopamine production by unaffected surrounding neurons, and upregulation of dopamine receptors in the target neurons in the striatum [23].

Diagnosis & treatment
PD is a clinical diagnosis based on the overall presentation of a patient. Preliminary premotor symptoms, motor symptoms, and cognitive and psychiatric symptoms have to be recognized. Premotor

symptoms include anosmia, rapid eye movement behaviour disorder, and depression. Motor symptoms include rigidity, bradykinesia, tremor, and balance and walking problems [3]. The altered gait pattern is characterised by a small stride length, reduced arm swing, stooped posture, shuffling feet, difficulties with turning, and freezing of gait. Autonomic dysfunction like constipation, hypotension and urinary urgency, as well as pain and cognitive decline can develop during later stages of the disease [1].

Additional research for diagnosis includes MRI, mainly to reject other causes of Parkinsonian syndromes [27]. Optionally, a FP-CIT-SPECT-scan ("DAT scan") can be performed in case a clinical diagnosis of a hypokinetic movement disorder cannot be made [27].

Treatments aim to suppress symptoms. The main treatment consists of pharmacotherapy with levodopa, which can be combined with or substituted by dopamine agonists. Besides pharmacotherapy, also physiotherapy, occupational therapy, and symptom specific treatments can be offered. Advanced therapy in patients with advanced PD includes deep brain stimulation, apomorphine, or levodopa/carbidopa intestinal gel [27].

Freezing of gait

As stated before, FOG is a symptom of PD usually defined as a "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk". Patients often describe FOG as the feeling of their feet being glued to the floor. Three FOG patterns are recognised: shuffling forward (very short, shuffling steps), trembling in place (alternating tremor of the legs), and complete akinesia (no movement of the limbs or trunk, uncommon). See also figure 3. FOG is often preceded by festination, the tendency to move forward with an increasing cadence and a decreasing step length. FOG can be asymmetrical and affect mainly one foot or turning in one direction. Episodes usually last a couple of seconds, although duration varies from less than a second to more than 30 seconds [4]. FOG is associated with the severity of PD and the duration of levodopa treatment, but can be present early in the disease or in untreated patients [3].

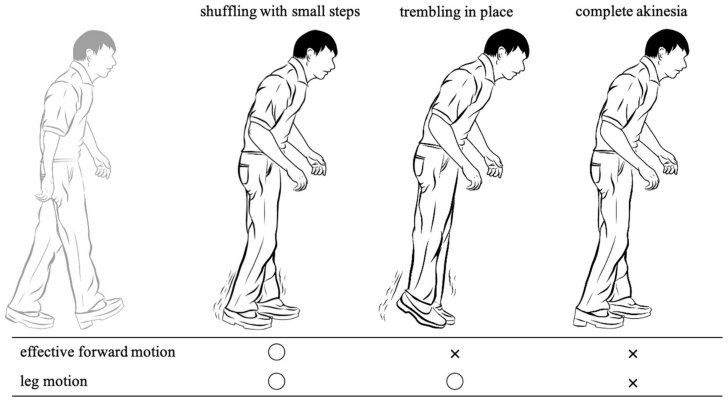


Figure 3: Three phenotypes of FOG according to leg movements. Picture from Kondo et al. [28].

Situations that may provoke FOG include gait initiation, turning, approaching a destination, and passing through narrow passages. Also environmental influences, emotional and cognitive situations can affect FOG. Approaching doorways, crowded places, dual-tasking, distractions, and time pressure enhance the likelihood that FOG occurs. Contrary, excitement, climbing stairs, and cueing strategies diminish the likelihood of FOG. Conditions that distract the patient from walking aggravate FOG, while paying attention to stepping alleviates FOG. When paying attention to walking, the impaired subcortical control of gait is taken over by the cortical pathway. In this way, the automatic control of walking is changed to goal-directed control [4].

2.2 Technical background

FOG detection based on IMU data is a type of multivariate time series classification. Classifiers are functions that use the input data to predict a probability distribution over the class variable values. FOG detection considers a binary classification problem: FOG or no FOG. The input is multivariate using several three-dimensional sensors, which means the discriminatory features may be in the interactions between dimensions [20]. In deep learning, the model performs feature extraction itself and then uses this features to classify the data, see figure 4. Classifiers, especially deep learning classifiers, require a large data set to learn to properly differentiate between classes. In this study, three deep learning classifiers are trained and tested: a self-developed CNN, MiniRocket, and InceptionTime. The CNN is a sequential model and has a relatively simple architecture. In MiniRocket, the data is transformed using a small, fixed set of convolution kernels and passed to a linear classifier [18]. InceptionTime is an ensemble of CNNs [19].

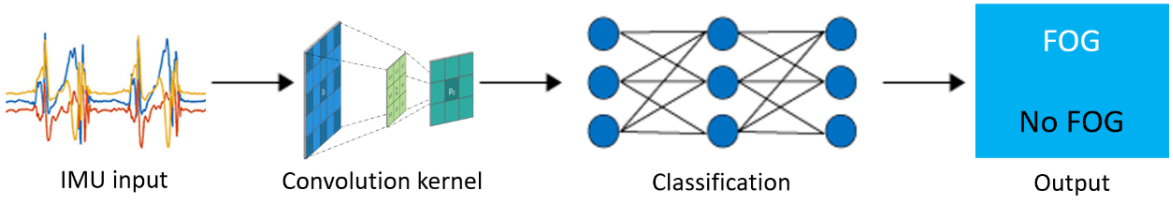


Figure 4: FOG classification using CNNs as deep learning method.

Convolutional neural network
A CNN is a neural network that consists of interconnected convolution layers of nodes, inspired by the structure of neurons in the brain. CNNs have proven to be successful at computer vision tasks like recognizing objects or faces. Although CNNs are designed for image data, they can also be used for time series data [11]. The CNN consists of several layers:

- Convolution layer: performs convolutions using convolution filters to produce feature maps. In order to perform the convolution, the filter is moved to every possible position in the feature map. Matrix multiplication is performed at every position and the results are added. Filter weights are determined during training. The feature map contains the features extracted from the input data [14];
- ReLU layer: the convolution layer is followed by a rectified linear activation unit (ReLU) function. The ReLU function sets negative values from the filtered matrix to zero in order to make sure that only neurons contributing to classification are activated [14];
- Batch normalization: one challenge of training neural networks is a changing distribution of the inputs to layers deep in the neural network, since parameters of the previous layers change. This may cause the model to keep chasing a moving target. Batch normalization is a technique to stabilize the learning process and reduce the number of required training epochs. The inputs to a layer are standardized for each mini-batch, which allows to use higher learning rates and make the algorithm less sensitive to initialization [29];
- Pooling layer: locally summarizes the input in order to make the output invariant to small translations in the input. In addition, the size of the activations that are fed to the next layer is reduced. Therewith, memory footprint is reduced and the overall computational efficiency improved. Max pooling, taking the maximum value of the neighborhood, is often used as pooling technique [14];
- Dropout layer: randomly and uniformly turns off non-output nodes, with a certain probability. It is a computationally cheap and effective regularization method to reduce overfitting and improve generalization [30];

- Softmax layer: the activation function in the last (dense) layer is softmax, assigning a probability to each class. The probabilities must sum up to 1. It is a type of logistic regression [31].

MiniRocket
The random convolutional kernel transform (Rocket) uses a large number of random convolution kernels in combination with a linear classifier. Input data is transformed using convolutional neural networks. Proportion of positive values (PPV) pooling is applied to the convolutional output to produce two features per kernel. This transformation results in a feature map. Logistic regression is applied to classify the data. Ruiz et al. recommended Rocket as the default choice for multivariate time series classification problems, because of the highest performance ranking and being the fastest classifier [20]. Rocket's successor MiniRocket maintains essentially the same accuracy, but is up to 75 times faster. The difference of MiniRocket compared to Rocket is the use of a small, fixed set of kernels, thereby using a mostly-deterministic approach. The kernels have only a length of 9, and all these weights are restricted to have two different values (-1 and 2). Therefore, MiniRocket should be used as the default variant of Rocket [18]. The architecture of MiniRocket is shown in figure 5.

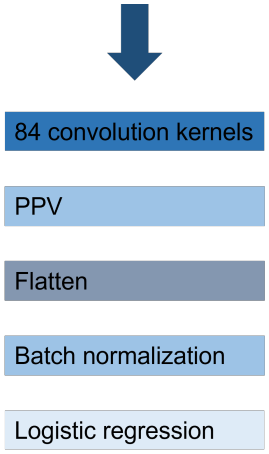


Figure 5: Architecture of MiniRocket. PPV = proportion of positive values, a pooling layer.

InceptionTime
InceptionTime is an ensemble of five deep convolutional neural network models developed for time series classification. The network outperforms the previous state-of-the-art deep learning model (HIVE-COTE) both in accuracy and scalability. Therefore, InceptionTime is recommended by Ruiz et al. as a deep learning method for time series classification [20]. InceptionTime consists of five inception networks. Each network is build with a series of inception modules followed by a global average pooling layer and a dense layer with a softmax activation function, see figure 6. An inception module consists of several layers (see figure 7):

- Bottleneck layer: the dimensionality or depth of the input data is reduced in this layer. Reducing dimensionality decreases the computational cost and the number of parameters, which results in faster training and improved generalization [32].

- Convolution layer: the output of the bottleneck layers enters three convolutional layers. These layers differ in kernel size, namely 10, 20 and 40 [32]. In this way, the network can automatically extract relevant features from both short and long time series [33].

- Max pooling layer: parallel to the bottleneck and convolution layer, a max pooling layer is added. The layer makes the model invariant to small perturbations. The max pooling layer is followed by a bottleneck layer [19].

- Depth concatenation layer: in the last layer of the inception module, the outputs of the four convolutional layers are concatenated along the depth dimension [32].

Furthermore, at every third inception module a residual connection is added. A residual or skip connection is a connection to pass data to latter parts of the neural network by skipping some layers, as showed in figure 8. The residual connection applies identity mapping to the input, and then performs element-wise addition of the input and the output of the inception modules. Residual connections enable deep neural networks to converge much more easily by avoiding exploding or vanishing gradient problems [34].

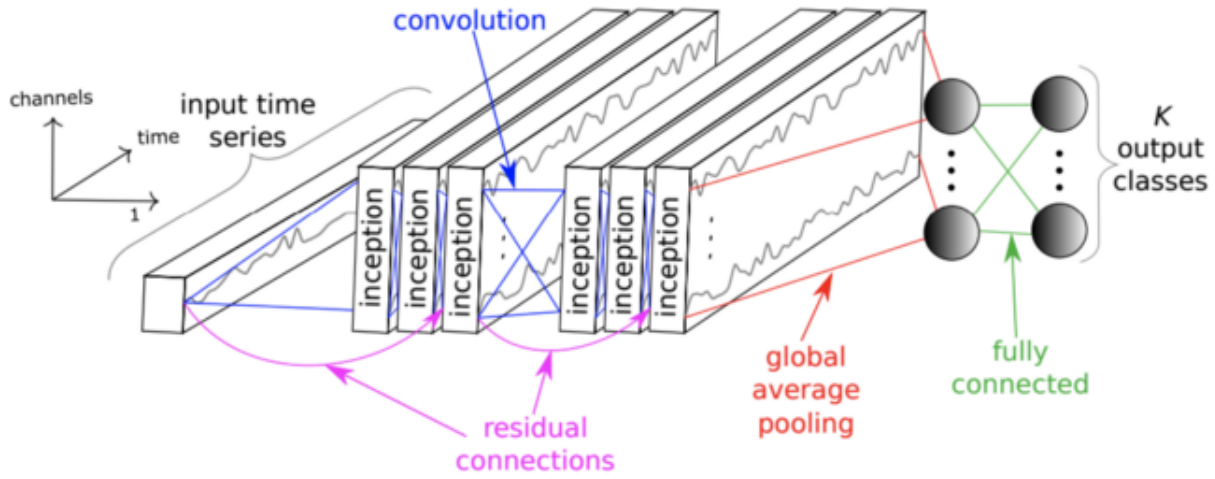


Figure 6: The Inception network for time series classification. Picture from Ismail-Fawaz et al. [19].

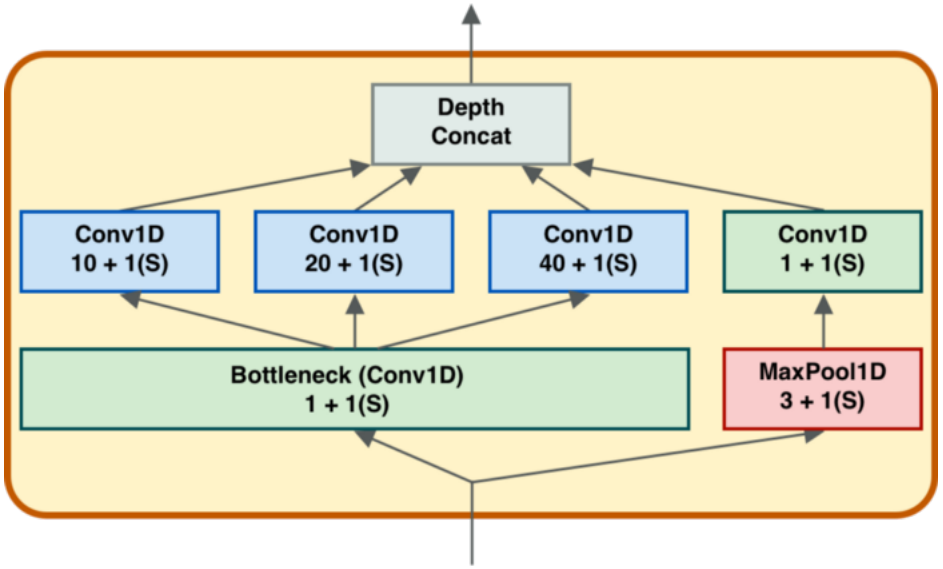


Figure 7: The inception module of InceptionTime. The first number in the boxes indicates the kernel size while the second indicates the size of the stride. "(S)" specifies the type of padding, i.e. "same". Picture from V. Stylianou [32].

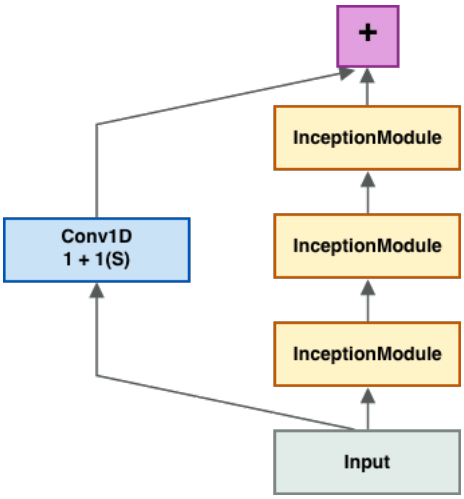


Figure 8: A residual connection in the inception network. Picture from V. Stylianou [32].

Adam optimizer

Adam is an optimization algorithm that can be used to update network weights during training. It is an extension to the stochastic gradient descent method. Stochastic gradient descent uses a fixed learning rate for all weight updates during training [35]. In Adam, adaptive moment estimation, adaptive learning rates for different parameters are used. The algorithm updates exponential moving averages of the gradient and the squared gradient. The moving averages are estimates of the mean and the uncentered variance of the gradient. The decay rates of these moving averages are controlled by two hyperparameters [36]. Adam is computationally efficient, requires little memory, and the default configuration parameters do well on most problems [35].

Cross-validation

Cross-validation is a statistical method to test the performance of a model and to find the optimal model. Several techniques are available, for example k-fold cross-validation, leave one group out cross-validation, time series cross-validation and nested cross-validation [37]. In this study, leave one group out cross-validation is used. The training+validation data is divided into five groups. All data of one patient is assigned to the same group. The different groups are balanced with regard to freezers/non-freezers and study, also called stratification. In this way, each subset contains the same skewed class distribution [38]. The network is trained from scratch five times, each time with a different validation set. See figure 9. The average performance of these five trained networks is thought to be representative for the performance of the model in general [39].

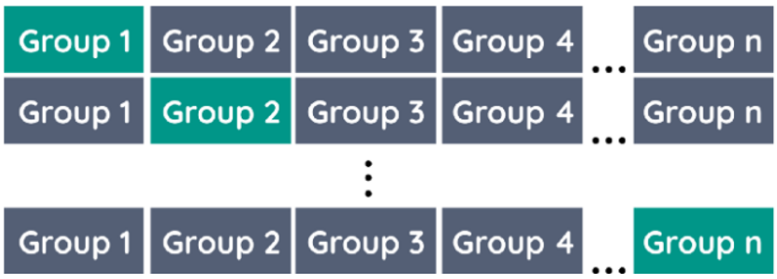


Figure 9: Leave one group out cross-validation. Each row represents one fold, with the highlighted group being the validation set. Picture from M. Abdelrazek [39].

3 Methods

In order to answer the research question, three classification models were examined. A dataset of 71 patients was used and cross-validation was applied. Next, the IMU sensor type and sensor locations were evaluated. The data is preprocessed using Matlab R2021b (The Mathworks, Natick, Massachusetts, USA) and Python version 3.9, and the algorithm was developed using Python version 3.9 (Python Software Foundation, Fredericksburg, Virginia, USA). The classification models were trained on a NVIDIA GeForce GTX 1080 Ti.

3.1 Data acquisition

In this study, the data from four previous studies was combined to create a large dataset: the Vibrating Socks study [40], the Pedal study [41], the Cinoptics study [42], and the Hololens study [43]. The medical ethical committee gave its approval to use the datasets for this non WMO research (Dutch Medical Research with Human Subjects Law). The study was assessed by the MEC-U and registered under W22.089. Furthermore, the committee gave its approval to link overlapping patients, who took part in multiple studies, to each other. This information is important to divide the dataset into a training, validation and test set with different participants. All experiments were recorded for video annotation of FOG by two experienced clinicians reaching consensus. IMU measurements were performed using the MVN Awinda motion capture system (XSens 3D motion tracking technology, Enschede, The Netherlands). The Xsens software to analyze the IMU data was MVN studio v4.2.0 for the Cinoptics study, MVN studio v4.4 for the Pedal study and the Hololens study, and MVN analyze 2020.0.1 for the Vibrating Socks study (XSens 3D motion tracking technology, Enschede, The Netherlands). The calibrated acceleration data and angular velocity data was provided in the earth-referenced local frame [44]. A short description of the studies is given below. An overview of the gait tasks and cueing conditions per study can be found in table 1.

Vibrating Socks study
Patients with PD and a recent history of disabling or regular FOG were included in the Vibrating Socks study, resulting in 32 participants. The data of 27 patients was included in this freeze detection study. Two measurements per patient were conducted, one while on dopaminergic medication and one while off dopaminergic medication. Patients were asked to perform three walking tasks, which were repeated four times with different cueing conditions [40].

Pedal study
Patients with PD and a recent history of disabling or regular FOG were included in the Pedal study, resulting in twenty participants. The same number of healthy controls were included, but left out of the dataset in this study. Measurements were performed while patients were off dopaminergic medication. Patients were asked to walk 30 meters straight in a corridor with two narrow passages at 10 and 20 meters. When patients reached the end of the corridor, they were requested to walk a wide turn and return. Multiple trials lasting 30 seconds each were performed directly after each other. The experiment was performed with different cueing conditions and cognitive conditions. The adjusted auditory stroop task (AAS) was applied to increase cognitive load, aiming to induce FOG. Besides the walking task, also a pedaling experiment with virtual reality was performed [41]. This part of the Pedal study was not included in the dataset of this study.

Cinoptics study
Patients with PD and a recent history of disabling or regular FOG were included in the Cinoptics study, resulting in twenty-five participants. Measurements were performed while patients were end-of-dose levodopa. Patients were asked to perform three walking tasks, with five different cueing conditions. A narrow passage was located halfway down the trajectory. The whole experiment was performed twice [42].

Hololens study
Patients with PD and a recent history of disabling or regular FOG were included in the Hololens study, resulting in eighteen participants. Patients were asked to perform a turning task under four different conditions. For each cueing condition, patients performed 15 trials of 180 degree turns on the spot [43].

Table 1: Gait tasks and cueing conditions of the four datasets.

<table>
<tr><th></th><th>Vibrating Socks</th><th>Pedal</th><th>Cinoptics</th><th>Hololens</th></tr>
<tr><td rowspan="3">Walking task</td><td>Walking 10 meter straight.</td><td rowspan="3">Walking straight with two narrow passages for 30 seconds.</td><td>Walking 15 meter straight.</td><td rowspan="3">180 degree turns to the left and to the right.</td></tr>
<tr><td>Gait trajectory consisting of two turns (left and right) and a narrow passage.</td><td>Walking 15 meter straight with stop and start commands at three randomized moments.</td></tr>
<tr><td>360 degree turns to the left and to the right.</td><td>Walking 15 meter straight with 360 degree turns at three randomized moments.</td></tr>
<tr><td rowspan="4">Cueing</td><td>Tactile cueing in a closed-loop manner.</td><td>Visual cueing using white transverse bars.</td><td>Visual cueing with augmented 3D bars.</td><td>Auditory cueing.</td></tr>
<tr><td>Tactile cueing in an open-loop manner.</td><td>Visual cueing using white transverse bars and AAS.</td><td>Visual cueing with augmented 3D staircase.</td><td>Visual cueing in augmented reality.</td></tr>
<tr><td>Auditory cueing.</td><td>No cueing.</td><td>Visual cueing with conservative bars on the floor.</td><td>Both auditory and visual cueing.</td></tr>
<tr><td rowspan="2">No cueing.</td><td rowspan="2">No cueing and AAS.</td><td>Auditory cueing.</td><td rowspan="2">No cueing.</td></tr>
<tr><td>No cueing.</td></tr>
</table>

3.2 Data preprocessing

The IMU data was synchronised with the labeled video data. Artefacts were detected and removed. Artefacts were defined as acceleration >100 m/s^2 or angular velocity > 20°/s. The IMU data was filtered with a zero phase third order bandpass Butterworth filter. The cut-off frequencies of 0.3 Hz and 15 Hz were chosen to remove the offset and to enable a full spectral analysis on the locomotor band of 0.5-3 Hz and freezing band of 3-8 Hz, while higher harmonics were removed [3, 16, 45]. The data was resampled to 60 Hz in order to ensure a uniform sampling frequency. The dataset was split in a training + validation and testing group. The data within these subsets was balanced regarding the presence of FOG and the different studies. The data was scaled per channel between -1 and 1, based on the minimum and maximum value in the whole training set. The data was windowed using sliding windows of 2 seconds every 0.5 seconds, corresponding to 75% overlap [12]. Windows without any FOG samples were labeled as no FOG, windows containing ≥25% FOG were labeled as FOG. Windows with >0% and <25% FOG were discarded from the training and validation set and from the test set. A similar approach was used by Sigcha et al. and Camps et al., however the percentage

of FOG was lowered to 25% in this study in order to retain more FOG-labelled windows in the data set [12, 46]. Twenty-five percent of FOG was hypothesized to be high enough to train the network with the characteristics of FOG episodes, while not loosing too much FOG data. In this way, FOG episodes of at least 0.5 seconds were taken into account.

3.3 Network architecture

Three classification models were evaluated on the IMU data of patients with PD. First of all, a basic CNN inspired by previous FOG detection networks is evaluated. Second, a fast and top-performing time series classification model named MiniRocket is examined. Lastly, the state-of-the-art deep learning classifier for time series data called InceptionTime is evaluated on IMU data. The input data for the comparison between the models was the acceleration data of the lower legs and feet. All models were optimized using Adam. An overview of the hyperparameters can be found in table 2.

Table 2: Hyperparameters of the three deep learning models.

Hyperparameter	**CNN**	**MiniRocket**	**InceptionTime**
Initial learning rate	0.001	0.001	0.001
Loss function	Binary cross entropy	Logistic regression	Binary cross entropy
Number of epochs	30	10	20
Kernel size	6	9	2, 4, 8
Class weight FOG	Balanced	Determined by algorithm	Balanced

Convolutional neural network
The model consists of three convolutional layers with kernel size 6. The kernel size was optimized to gain the highest area under the receiver operating characteristic (AUC-ROC). Each convolution layer with ReLU function is followed by batch normalization, max pooling, and a dropout layer. After three of these blocks, a flatten layer generates one dimensional input for the two dense layers performing actual classification of the data. The output of these dense layers is a probability distribution over the class variables. The architecture can be found in figure 10.

Figure 10: Architecture of the proposed CNN.

The model was implemented using Tensorflow in Python. The loss function is binary cross entropy. In order to overcome the unbalance of the dataset, a balancing class weight (calculated using the percentage of FOG) was used to penalize the algorithm harder for falsely classifying freezing windows. The model was trained for 100 epochs to check the training process, and thereafter set to 30 epochs.

MiniRocket
MiniRocket was implemented using the PyTorch version of the time series for artificial intelligence (tsai) package in Python. As recommended by Dempster et al. for datasets of >10 000 samples, a logistic regressor was applied. More information about this model can be found in chapter 2 and in Dempster et al. 2021 [18]. The learning rate is halved if the validation loss does not improve after 50 updates, and training is stopped after 100 updates without improvement of the validation loss. The

class weight is determined by the algorithm itself.

Inception Time
Inception Time was implemented using Tensorflow and the classification algorithm provided on Github by Ismail Fawaz et al. More information about this model can be found in chapter 2 and in Ismail Fawaz et al. [19]. Instead of the default kernel size of 10, 20 and 40, a smaller kernel size of 2, 4 and 8 was used to decrease the number of parameters in the network and thus prevent overfitting [19]. The loss function is binary cross entropy, a popular method for binary classification taking the log of the probabilities [47]. In order to overcome the unbalance of the dataset, a balancing class weight was used to penalize the algorithm harder for falsely classifying freezing windows. The model was trained for 100 epochs to check the training process, and thereafter set to 20 epochs.

3.4 Sensor evaluation

Accelerometer data of the lower legs and feet was selected from the IMU data as input to the three models. However, IMU data also includes gyroscope data that may provide additional information about the presence of FOG. In this study, we evaluated whether adding gyroscope data enhances the model performance of the model with the highest AUC-ROC. Furthermore, body locations of the IMU sensors were evaluated. A trade off between increase in model performance and practical applicability has to be made. Sensor combinations that were studied include:

- Upper legs + lower legs + feet;
- Lower legs + feet;
- Feet.

3.5 Model validation and analysis

To test if the three models were implemented correctly, the models were trained and evaluated on the Wafer dataset from the UCR time series classification repository. The UCR repository is an archive with 85 time series datasets. The data is z-normalized to remove offset and scaling and has a predefined train/test split. The archive contains among others IMU data, but no gait tasks [48].

Cross-validation was implemented in order to estimate an unbiased model performance [39]. Five-fold leave one group out stratified cross-validation was applied. All data of one patient was assigned to the same fold, taking into account participants that took part in multiple studies. Stratification of study and freezers/non freezers was applied to maintain the same class distribution in each subset [38]. The average performance of the five validation folds was used to compare several models.

The models were evaluated for various discrimination thresholds using a receiver operating characteristic (ROC) curve. The best model was chosen based on the AUC-ROC. The final model was evaluated on the hold out test set. In this way, the optimized model could be checked for overfitting and a fair performance was calculated. Besides the ROC curve, the sensitivity and specificity with Clopper-Pearson 95% confidence interval were calculated based on the confusion chart.

4 Results

In total, 71 unique participants were included in this study. Fourteen of them participated in multiple studies, all data of one patient was assigned to either the training, validation or test set. Fifty-seven patients were assigned to the training + validation set and fourteen to the test set. Clinical characteristics of the participants can be found in table 3. The clinical characteristics could not be determined separately for the training + validation and test set, since the individual patient characteristics were unknown in this study. Characteristics about the number of patients, size of the datasets and percentage of FOG of the training + validation and test group can be found in table 4.

Table 3: Clinical characteristics of the participants per study.

	Vibrating Socks	**Pedal**	**Cinoptics**	**Hololens**
	Median (25th - 75th percentile)			Median (IQR)
No. participants	32	20	25	16
Age (years)	66 (60 - 74)	70.5 (63.5 - 73)	72 (65 - 79)	69 (13)
Gender (% male)	87.1	85	76	81
Disease duration (years)	11 (5 - 14)	11 (7.5 - 16)	11 (3 - 20)	10 (9)
Years since FOG (years)		4 (2.5 - 6.5)	2 (0.25 - 12)	4 (9)
Hoehn and Yahr score	2 (2 - 3)	2 (2 - 3)	2 (2 - 3)	2 (1)
MDS-UPDRS III	ON: 38 (29 - 46) OFF: 51 (47 - 62)	39.5 (31.5 - 47.5)	34 (10 - 61)	38 (17)
N-FOGQ	FOGQ: 15 (13 - 18)	21 (16 - 25)	18 (8 - 28)	18 (7)
MMSE	28 (26 - 30)	29 (27 - 30)	28 (19 - 30)	29 (2)
FAB	17 (14 - 18)	16 (15 - 17)	14 (5 - 26)	17 (2)

MDS-UPDRS III, Movement Disorders Society Unified Parkinson's Disease Rating Scale part III; N-FOGQ, New Freezing of Gait Questionnaire (range 0-28); FOGQ, Freezing of Gait Questionnaire (range 0-24); MMSE, mini-mental state examination (range 0-30); FAB, Frontal Assessment Battery (range 0-18); IQR, interquartile range; ON, 'on' dopaminergic state; OFF, 'off' dopaminergic state.

Table 4: Characteristics of training + validation group and test group.

Characteristics	**Training + validation set**	**Test set**
No. patients	57	14
- Vibrating Socks study	21	6
- Pedal study	17	3
- Cinoptics study	21	4
- Hololens study	15	3
No. patients in multiple studies	12	2
No. samples	131 372	40 938
- Vibrating Socks study	71 946	29 648
- Pedal study	22 875	4 137
- Cinoptics study	24 555	4 786
- Hololens study	12 996	2 367
No. samples with 0-25% FOG	0	0
No. samples with 25-100% FOG	10 590	4 978
Percentage of (25-100%) FOG	8.1	12.2
- Vibrating Socks study	4.8	9.5
- Pedal study	1.1	12.9
- Cinoptics study	11.2	0.2
- Hololens study	31.9	78.0

Prior to training the models with the IMU data, the models were tested using the Wafer dataset of the UCR archive [48]. The performance of MiniRocket and InceptionTime on this dataset was comparable with the performance of the authors of these algorithms [18, 19]. The performance of the CNN was comparable with the performance of the other two models. Results can be found in appendix A.

4.1 Neural network evaluation

The CNN, MiniRocket and InceptionTime were trained with the acceleration data of the lower legs and feet. The ROC curves and model loss of the CNN, MiniRocket and InceptionTime can be found in figure 11.

Convolutional neural network
The ROC curve of the CNN in figure 11a shows the performance after 30 epochs with a mean AUC-ROC of 0.82 ±0.03. The performance difference in between the folds is limited, indicated by the standard deviation of the mean AUC-ROC. The number of epochs was determined based on the model loss in figure 11b. The validation loss varies around 0.6 during the training process and does not decrease further, so the training is stopped after 30 epochs. The validation loss does not seem to be affected by the training process. The training loss is smaller than 0.1 from the beginning, and decreases further over time. The training loss was expected to start at a higher level.

MiniRocket
The ROC curve of MiniRocket in figure 11c shows the performance with a mean AUC-ROC of 0.72 ±0.09. The AUC-ROC of the individual folds varies from 0.61 to 0.86, both outliers compared to the mean AUC-ROC and its standard deviation. The model loss could not be displayed for this algorithm.

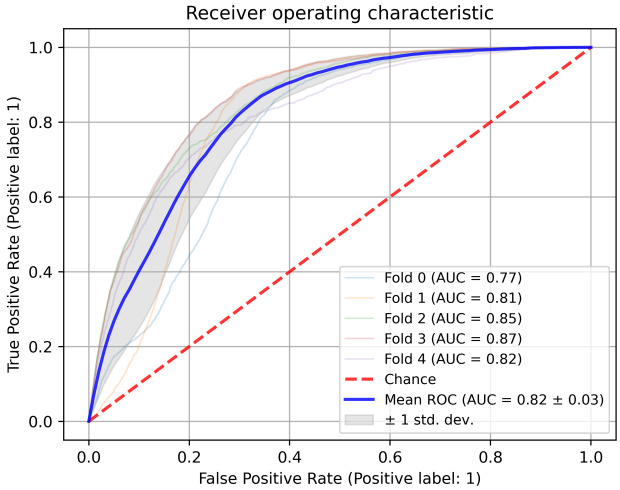
InceptionTime
Figure 11d shows the ROC curve of InceptionTime after 20 epochs with a mean AUC-ROC of 0.78 ±0.05. The variance in AUC-ROC is larger than for the CNN, but smaller than for MiniRocket. Figure 11e shows that the training loss is below 0.1 from the beginning and decreases further, while the validation loss increases during the training process indicating overfitting. Therefore, the kernel size was reduced and the number of epochs was limited to twenty.

4.2 Sensor evaluation

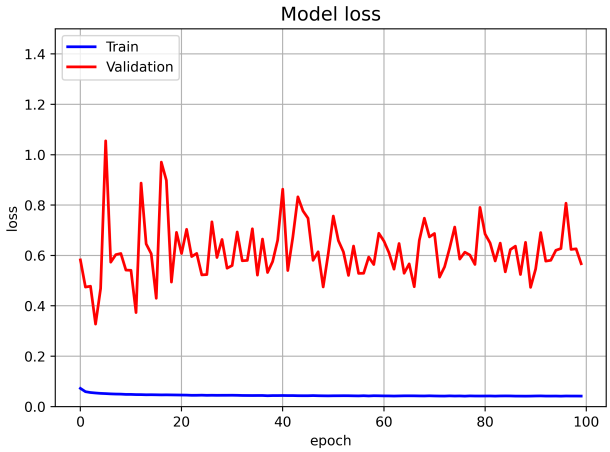
The sensor evaluation was performed on the CNN, since this model showed the highest AUC-ROC. The ROC curves of the CNN trained on 1) upper legs, lower legs, feet; 2) lower legs, feet; and 3) feet; can be found in figure 12. The corresponding AUC-ROCs are 1) 0.81 ±0.04; 2) 0.82 ±0.03; and 3) 0.81 ±0.04. Because of technical memory issues, the added value of gyroscope data could not be evaluated. Although the mean AUC-ROC values are almost the same, the input data with the highest AUC-ROC was selected for the final model, namely the acceleration data of the lower legs and feet.

4.3 Final model - evaluation on test set

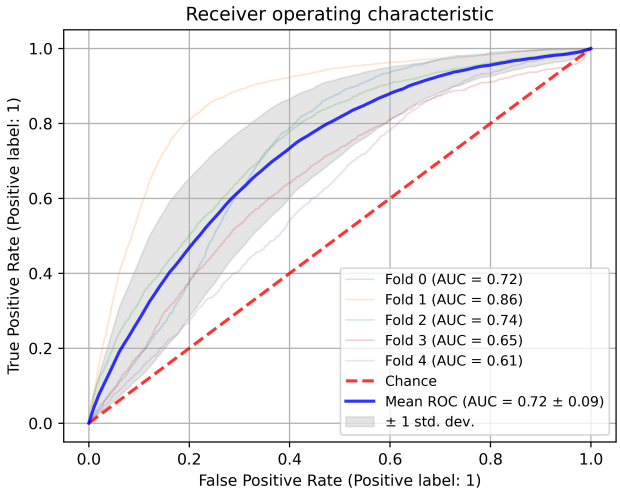
The final model is the CNN trained with acceleration data of the lower legs and feet. The ROC curve on the test set with an AUC-ROC of 0.72 is shown in figure 13a. With a threshold of 0.01, a sensitivity of 73.7% (72.5 - 75.0%) and specificity of 60.8% (60.3 - 61.3%) was reached. The threshold is based on the ROC curve of the test set, with the highest possible sensitivity while the specificity is above 60%. The sensitivity and specificity were deduced from the confusion chart in figure 13b.



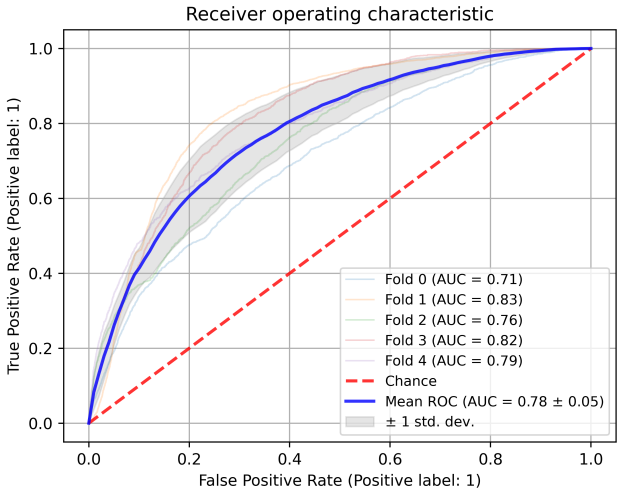
(a) ROC curve of the five folds and the mean ROC after training the CNN for 30 epochs.



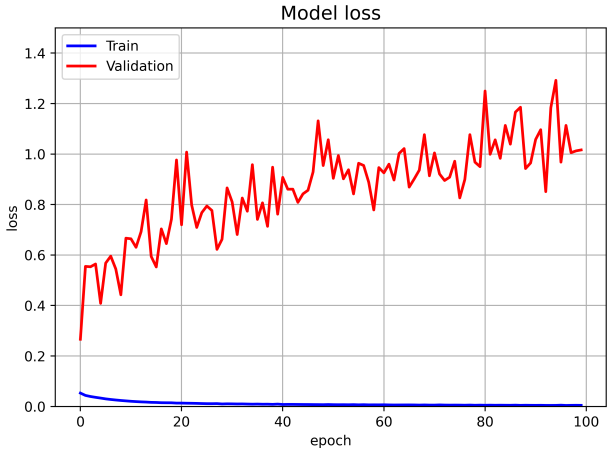
(b) Model loss of the first fold in training the CNN for 100 epochs.



(c) ROC curve of the five folds and the mean ROC after training MiniRocket.

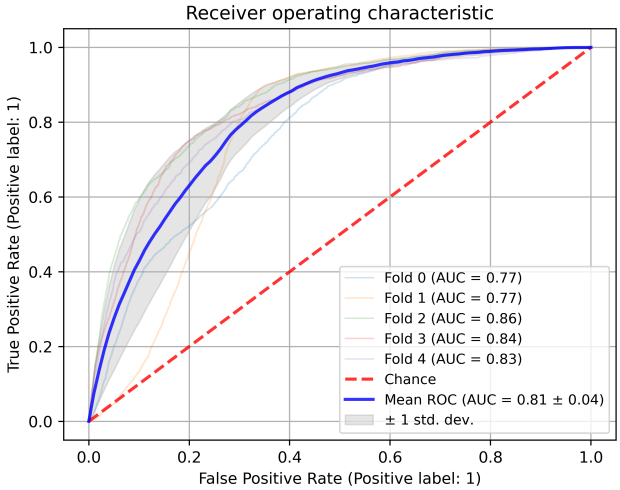


(d) ROC curve of the five folds and the mean ROC after training InceptionTime for 20 epochs.

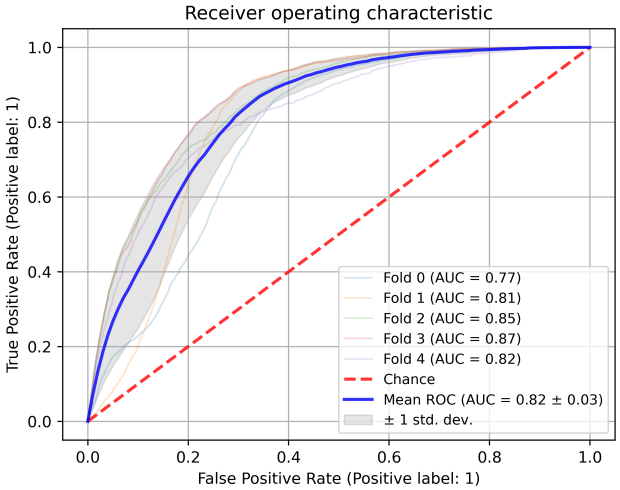


(e) Model loss of the first fold in training InceptionTime for 100 epochs.

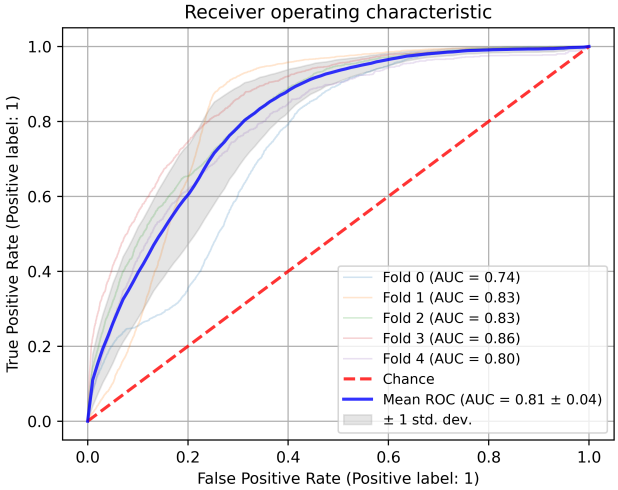
Figure 11: ROC curve and model loss of the three detection algorithms.



(a) ROC curve of the CNN trained on the acceleration data of the upper legs, lower legs and feet.

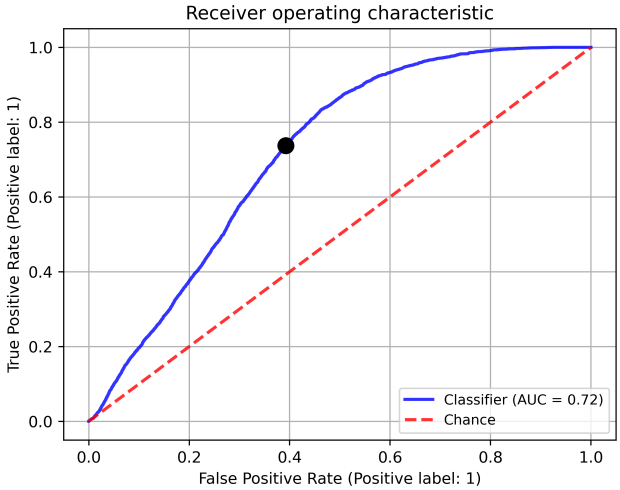


(b) ROC curve of the CNN trained on the acceleration data of the lower legs and feet.

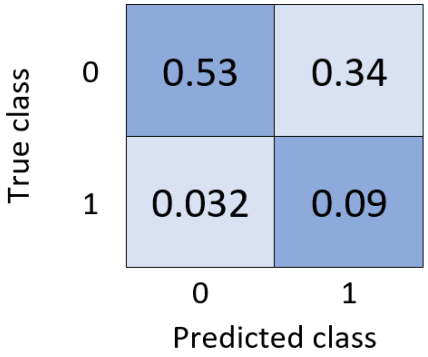


(c) ROC curve of the CNN trained on the acceleration data of the feet.

Figure 12: ROC curves of the CNN trained on three combinations of acceleration sensors.



(a) ROC curve of the final model. The black dot represents the threshold which was set at 0.01.



(b) Normalized confusion chart of the final model.

Figure 13: Performance of the CNN evaluated on the test set with acceleration data of the lower legs and feet as input.

5 Discussion

In this study, three classification models, namely a CNN, MiniRocket and InceptionTime, were implemented for the detection of FOG based on IMU data. The CNN trained on acceleration data of the lower legs and feet showed the highest AUC-ROC. When evaluated on the test set, an AUC-ROC of 0.72, sensitivity of 73.7% (72.5 - 75.0%) and specificity of 60.8% (60.3 - 61.3%) was reached.

5.1 Interpretation of the results

Comparing the CNN, MiniRocket and InceptionTime with accelerometer data of the lower legs and feet of 71 PD patients as input, the CNN turned out to reach the highest mean AUC-ROC on the validation set (0.84 $\pm$0.04). For the CNN, the sensors of the lower legs and feet resulted in the highest performance, although the difference with other sensor combinations was only 0.01 in AUC-ROC. The added value of gyroscope data could not be evaluated. An AUC-ROC of 0.72 was reached on the test set. The decrease in AUC-ROC on the test set indicates that the model might be overfitting on the training + validation set.

As described in the hypothesis, MiniRocket was expected to show the best performance in this FOG detection study. However, the AUC-ROC was 0.10 smaller than the AUC-ROC of the CNN. The variance in the ROC curves of the five folds, figure 11c, is higher than for the CNN and InceptionTime, which might indicate overfitting. Therefore, the model loss should be inspected. The number of epochs was optimized by the algorithm. Other hyperparameters and the windowing should be optimized. Despite promising results on time series datasets including IMU data from the UCR archive [18, 20, 48], the architecture seems less effective on IMU data for FOG detection.

As hypothesized because of the complex architecture, InceptionTime showed severe overfitting. Therefore, the kernel size and number of epochs was reduced. More techniques to reduce overfitting are required. First, the AdamW optimizer should be implemented to improve regularization in Adam. This optimizer decouples the weight decay from the gradient-based update and this improves generalization [49]. Second, drop-out layers should be used, as done in the CNN. Next, further enlargement of the dataset is required to reach the optimal performance of this complex time series classification model in FOG detection.

Contrary to InceptionTime, at first sight the CNN suffers much less from overfitting, due to the simpler architecture. However, the model loss shows a different behaviour than expected. As can be seen from figure 11b, the training loss is very small from the beginning. The training loss decreases further, so the model is learning. The validation loss neither decreases, as expected during at least the first part of training, nor increases, indicating overfitting. The predictions of the validation set seem basically random. The code should be checked for bugs that may cause this unexpected behaviour. Another cause of the problem might be a different data distribution between the training and validation set. Deep learning models normally assume that the data is independent and identically distributed [50]. If that is not the case, these models fail to work. However, the data in the training, validation and test set was balanced with regard to freezers and study, and the results of the five folds were similar. Further research is needed to find the origin of the problem. The ROC curve after 30 epochs in figure 11a is comparable with the ROC curve after 100 epochs. This corresponds to the validation loss, which also varies around the same level. The number of epochs should be optimized further by automatically saving the model with the smallest validation loss.

The results of this study are in line with previous research in FOG detection using deep learning based on IMU data [11, 13]. In addition, previous pilot studies on part of the dataset showed similar results. For example, Opdam was able to train a CNN on the Cinoptics acceleration data with an AUC-ROC of 0.83 [21].

The methodology of this study is similar to other FOG detection studies [12, 14, 46]. What is remarkable are the data characteristics per study, see table 4. Even though the subsets are balanced with regard to freezers and studies, the percentage of FOG per study varies a lot. Reason for this is the big difference in amount of FOG per participant. In addition, over 50% of the dataset originates from the Vibrating Socks study and could therefore heavily influence the results. Further balancing of the data could help the training process and ensure a fair performance. Another setting that affects the performance is the FOG labelling with a threshold of 25% or 0.5 seconds. A high threshold improves the model performance, while a low threshold enables recognition of partial FOG windows and short FOG episodes. Thus lowering the threshold minimises the detection latency and improves detection robustness. However, this comes at the cost of the model performance [13].

5.2 Strengths and limitations

This study has several strenghts. First of all, the algorithms in this study are trained on a large dataset with IMU data of 71 PD patients with FOG. As stated before, deep learning requires a large dataset to prevent overfitting. However, the datasets used in previous research were limited in size (up to 21 patients) [11]. Recently, Shi et al. published a FOG detection study on 67 PD subjects. Shi et al. proposed a deep learning model and compared the model against the state-of-the-art deep learning models for FOG detection from Xia, Camps and Bikias [46, 51, 52]. They showed a decreased performance on the reconstructed models. Reason for this may be a different dataset, an increased heterogeneity in the data because of a larger dataset, differences in data collection, sensor data, or preprocessing [13]. These results emphasize the need for a large dataset for deep learning algorithms. Although the dataset in this study is much larger than most of the published studies on FOG detection, more data would improve the training process of the deep learning models and allow the use of more complicated architectures.

Second, the dataset includes gait tasks ranging from straight walking, turning and obstacle course. These different gait tasks make the data more representative for movements in real life. Third, all data was gathered using the hardware and software of Xsens, which ensures a consistent and reliable dataset. Different software versions of MVN studio and MVN analyze were used in different studies, which may result in minor differences in the calibrated and normalized data.

Fourth, cross-validation was applied to estimate an unbiased model performance. The data in the training, validation and test set was balanced with regard to the presence of FOG and the different studies. All data of one participant was assigned to the same subset. For patients who took part in multiple studies, the identification numbers of the different studies were linked to make sure that all data of one participant was assigned to the same subset.

Fifth, three promising classification algorithms were evaluated for the application of FOG detection. The CNN was inspired by the model of Abdallah et al. [14] and optimized for this binary FOG classification problem. The relatively simple architecture turns out to be effective on the dataset in this study. MiniRocket and InceptionTime are considered to be state-of-the-art models in time series classification, and have not been evaluated on IMU data for FOG detection before [18–20]. It can be concluded that these models require optimization specifically for IMU datasets containing FOG.

Finally, another strength of this study is the patient independent models, since these are easily applicable to cueing devices like the vibrating socks. Automatic FOG detection enables on-demand cueing, but also objective assessment of therapy using home monitoring.

Besides some strengths, this study has several limitations. The dataset in this study is a combination of four studies. Combining datasets may cause different artefacts, a different scale because of the

different software versions of Xsens, or a different gait pattern because of the provided cues [21]. In addition, the percentage of FOG varies a lot per study. Because of these differences in the datasets, the networks may learn to recognize which samples originate from which dataset. On the other hand, big artefacts were removed, and the different software versions of Xsens are thought to have minimal effects on the normalized data. Considering the variation in gait pattern and FOG frequency, the algorithm should recognize FOG regardless of the specific movement at that moment. After all, in reality the algorithm should also detect FOG regardless of the walking pattern. In order to exclude the possible negative effects of combining datasets, the performance per individual dataset and per gait task should be evaluated.

The second limitation is the calibration of the Xsens IMU data. As described in chapter 3, the data is calibrated in the earth local frame, so walking forward is presented opposite from walking backward [44]. This problem applies to the Vibrating Socks and Cinoptics study, and to some patients of the Pedal study. It would be favorable to use free acceleration, comparable to IMU sensors in cueing devices. To train the algorithm with more universal data, the part where patients walk back should be flipped.

Third, the algorithms are trained on FOG annotated by two experts per study. Although this is the gold standard, there is often discussion about the classification between experts [10]. Thus the assumed ground truth may be incorrect sometimes. In addition, FOG is known for different phenotypes and diverse presentation in between patients which complicates detection [28].

Finally, windows containing >0% and $\leq$ 25% FOG were removed from the training + validation and test set. The data is treated as binary, but for the windows at the beginning and end of a FOG episode this is not true. For that reason, windows with a small amount of FOG were removed to make the classification problem binary. In reality, the models would randomly assign these windows to a class. Therefore, the performance on windows with a small amount of FOG should be evaluated separately.

5.3 Future research

Besides optimization of the hyperparameters, preprocessing settings such as the window length and overlap should be varied. According to Tautan et al., the input data should be provided as accelerometer magnitude instead of the individual accelerometer axis with the following formula: $magnitude = \sqrt{x^2 + y^2 + z^2}$ with x, y, z the three accelerometer axes signals [53]. In addition, data augmentation should be explored to further enlarge the dataset. Data augmentation is widely applied in imaging, but has to be used with care for time series data [54]. Therefore, the models should be trained with and without data augmentation. Examples of applicable data augmentation are switching the sensors of the left and right limb, or adding noise. Next to applying data augmentation, adding more real data would improve the training of the network. The network architecture of the CNN could also be changed trying to improve the performance. Further validation of the models should be done with another dataset for FOG detection of IMU data, to test the models on a different study and show the models' ability to generalize. A dataset from the catholic university of Leuven is available for this evaluation.

Considering the performance of the CNN, the model has potential to be implemented in on-demand cueing devices and home monitoring applications for objective FOG detection. For practical applicability and patient comfort in the vibrating socks, it is recommended to select the sensors solely at the ankles to the cost of a slightly lower performance. In order to make the detection algorithm patient specific, the threshold to determine the point on the ROC curve should be variable. When a patient experiences a lot of favor from the vibrating socks, a high sensitivity can be chosen; while a patient who thinks the vibrations are annoying, could choose a higher specificity.

6 Conclusion

The aim of this research was to develop a detection model that classifies FOG in PD patients based on IMU data. Four datasets with IMU data of 71 PD patients were combined and used as input for the models. Three neural networks, namely a CNN, MiniRocket and InceptionTime, were implemented and trained on this dataset. The CNN trained on acceleration data of the lower legs and feet performed best, with an AUC-ROC of 0.72, sensitivity of 73.7% (72.5 - 75.0%) and specificity of 60.8% (60.3 - 61.3%) on the test set. The model has potential to be implemented in on-demand cueing devices and home monitoring applications for objective FOG detection. Therefore, further research is needed. The unexpected behaviour of the model loss should be declared and solved. Furthermore, the hyperparameters should be optimized further to prevent overfitting, preprocessing settings should be varied to obtain the optimal data input, and data augmentation should be explored. Also adjustments in the network architecture may improve the performance.

References

[1] David K. Simon, Caroline M. Tanner, and Patrik Brundin. Parkinson Disease Epidemiology, Pathology, Genetics and Pathophysiology. *Clinics in geriatric medicine*, 36(1):1, feb 2020. doi: 10.1016/J.CGER.2019.08.002.

[2] E. Ray Dorsey, Todd Sherer, Michael S. Okun, and Bastiaan R. Bloem. The emerging evidence of the Parkinson pandemic, 2018. ISSN 1877718X.

[3] Steven T. Moore, Hamish G. MacDougall, and William G. Ondo. Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of Neuroscience Methods*, 167(2):340–348, jan 2008. ISSN 0165-0270. doi: 10.1016/J.JNEUMETH.2007.08.023.

[4] John G Nutt, Bastiaan R Bloem, Nir Giladi, Mark Hallett, Fay B Horak, and Alice Nieuwboer. Freezing of gait: moving forward on a mysterious clinical phenomenon. *The Lancet. Neurology*, 10(8):734, aug 2011. doi: 10.1016/S1474-4422(11)70143-0.

[5] Carola M. Koopman, Eric Lutters, Jorik Nonnekes, Bastiaan R. Bloem, Jeroen P.P. van Vugt, and Marleen C. Tjepkema-Cloostermans. Vibrating socks to improve gait in Parkinson's disease, dec 2019. ISSN 18735126.

[6] M.C. Tjepkema-Cloostermans, E.C. Klaver, B.R. Bloem, J. Nonnekes, and Jeroen P.P. van Vugt. Vibrating socks as a home-based tactile cueing device in Parkinson's disease - Research protocol. Technical report, Medisch Spectrum Twente, Enschede, 2021.

[7] R. Velik, U. Hoffmann, H. Zabaleta, J.F. Marti Masso, and T. Keller. The effect of visual cues on the number and duration of freezing episodes in Parkinson's patients. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2012:4656–4659, 2012. ISSN 2694-0604. doi: 10.1109/EMBC.2012.6347005.

[8] L.F.J. Nanning. Vibrating socks as a home-based tactile cueing device in Parkinson's Disease. Technical report, University of Twente, Enschede, 2021.

[9] Tiffany R. Morris, Catherine Cho, Valentina Dilda, James M. Shine, Sharon L. Naismith, Simon J.G. Lewis, and Steven T. Moore. Clinical assessment of freezing of gait in Parkinson's disease from computer-generated animation. *Gait Posture*, 38(2):326–329, jun 2013. ISSN 0966-6362. doi: 10.1016/J.GAITPOST.2012.12.011.

[10] Helena Cockx, Emilie Klaver, Marleen Tjepkema-Cloostermans, Richard van Wezel, and Jorik Nonnekes. The Gray Area of Freezing of Gait Annotation: A Guideline and Open-Source Practical Tool. *Movement Disorders Clinical Practice*, 2022. ISSN 2330-1619. doi: 10.1002/MDC3.13556.

[11] Scott Pardoel, Jonathan Kofman, Julie Nantel, and Edward D. Lemaire. Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: A review, dec 2019. ISSN 14248220.

[12] Luis Sigcha, Nélson Costa, Ignacio Pavón, Susana Costa, Pedro Arezes, Juan Manuel López, and Guillermo De Arcas. Deep Learning Approaches for Detecting Freezing of Gait in Parkinson's Disease Patients through On-Body Acceleration Sensors. *Sensors 2020, Vol. 20, Page 1895*, 20 (7):1895, mar 2020. ISSN 1424-8220. doi: 10.3390/S20071895.

[13] Bohan Shi, Arthur Tay, Wang Lok Au, Dawn May Leng Tan, Nicole Shuang Yu Chia, and Shih-Cheng Yen. Detection of Freezing of Gait Using Convolutional Neural Networks and Data From Lower Limb Motion Sensors. *IEEE Transactions on Biomedical Engineering*, 69(7):2256–2267, jul 2022. ISSN 0018-9294. doi: 10.1109/TBME.2022.3140258.

[14] Mostafa Abdallah, Ali Saad, and Mohamad Ayache. Freezing of Gait Detection: Deep Learning Approach. In *2019 International Arab Conference on Information Technology (ACIT)*, pages 259–261. IEEE, dec 2019. ISBN 978-1-7281-3010-1. doi: 10.1109/ACIT47987.2019.8991099.

[15] Rubén San-Segundo, Honorio Navarro-Hellín, Roque Torres-Sánchez, Jessica Hodgins, and Fernando de la Torre. Increasing Robustness in the Detection of Freezing of Gait in Parkinson's Disease. *Electronics 2019, Vol. 8, Page 119*, 8(2):119, jan 2019. ISSN 2079-9292. doi: 10.3390/ELECTRONICS8020119.

[16] Alexandra-Maria Tautan, Alexandra-Georgiana Andrei, and Bogdan Ionescu. Freezing of Gait Detection for Parkinson's Disease Patients using Accelerometer Data: Case Study. pages 1–4, dec 2020. doi: 10.1109/EHB50910.2020.9280223.

[17] J. Ten Broeke. Detecting Freezing of Gait in patients with Parkinson's disease. Technical report, Medisch Spectrum Twente, Enschede, 2021.

[18] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 248–257, aug 2021. doi: 10.1145/3447548.3467231.

[19] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, nov 2020. ISSN 1573756X. doi: 10.1007/S10618-020-00710-Y/FIGURES/23.

[20] Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 35(2):401–449, mar 2021. ISSN 1384-5810. doi: 10.1007/S10618-020-00727-3.

[21] Dylan Opdam. Investigating neural networks for Freezing of gait detection. Technical report, University of Twente, Enschede, 2019.

[22] Romée Snijders and Veerle Smit. *Compendium Geneeskunde*. Synopsis BV, 6 edition, 2016. ISBN 978-90-825709-1-5.

[23] M.J. Turlough Fitzgerald, G. Gruener, and E Mtui. *Clinical neuroanatomy and neuroscience*. Saunders, 6 edition, 2012. ISBN 978-0-7020-4042-9.

[24] Alexandru Andrusca. Basal ganglia: Gross anatomy and function, 2022. URL `https://www.kenhub.com/en/library/anatomy/basal-ganglia`.

[25] Alexxai V. Kravitz and Anatol C. Kreitzer. Striatal mechanisms underlying movement, reinforcement, and punishment. *Physiology*, 27(3):167–177, jun 2012. ISSN 15489213. doi: 10.1152/PHYSIOL.00004.2012/ASSET/IMAGES/LARGE/PHY0031201100003.JPEG.

[26] Jean Martin Beaulieu and Raul R. Gainetdinov. The physiology, signaling, and pharmacology of dopamine receptors. *Pharmacological reviews*, 63(1):182–217, mar 2011. ISSN 1521-0081. doi: 10.1124/PR.110.002642.

[27] Federatie Medisch Specialisten. Richtlijn Ziekte van Parkinson, 2020. URL `https://richtlijnendatabase.nl/richtlijn/ziekte_van_parkinson/startpagina_ziekte_van_parkinson.html`.

[28] Yuki Kondo, Katsuhiro Mizuno, Kyota Bando, Ippei Suzuki, Takuya Nakamura, Shusei Hashide, Hideki Kadone, and Kenji Suzuki. Measurement Accuracy of Freezing of Gait Scoring Based on Videos. *Frontiers in Human Neuroscience*, 0:309, may 2022. ISSN 1662-5161. doi: 10.3389/FNHUM.2022.828355.

[29] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, feb 2015. doi: 10.48550/arxiv.1502.03167.

[30] Jason Brownlee. A Gentle Introduction to Dropout for Regularizing Deep Neural Networks, 2018. URL `https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/`.

[31] Google Developers. Multi-Class Neural Networks: Softmax, 2022. URL `https://developers.google.com/machine-learning/crash-course/multi-class-neural-networks/softmax`.

[32] Vasilis Stylianou. Deep Learning for Time Series Classification (InceptionTime), 2020. URL `https://towardsdatascience.com/deep-learning-for-time-series-classification-inceptiontime-245703f422db`.

[33] Emma Amor. Deep dive into GoogLeNet Inception Network Architecture — ML Cheat Sheet, 2020. URL `https://medium.com/ml-cheat-sheet/deep-dive-into-the-google-inception-network-architecture-960f65272314`.

[34] Wanshun Wong. What is Residual Connection? A technique for training very deep neural networks, 2021. URL `https://towardsdatascience.com/what-is-residual-connection-efb07cab0d55`.

[35] Jason Brownlee. Gentle Introduction to the Adam Optimization Algorithm for Deep Learning, 2017. URL `https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/`.

[36] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2014. doi: 10.48550/arxiv.1412.6980.

[37] Maarten Grootendorst. Validating your Machine Learning Model — by Maarten Grootendorst — Towards Data Science, 2019. URL `https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7`.

[38] Jason Brownlee. How to Fix k-Fold Cross-Validation for Imbalanced Classification, 2020. URL `https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/`.

[39] Mohammed Abdelrazek. Model validation In ML (Part I). Definition: — by Mohamed Abdelrazek — Medium, 2021. URL `https://mohamedabdelrazek-14826.medium.com/model-validation-in-ml-part-i-ec827006de10`.

[40] Dr. M.C. Tjepkema, Prof. dr. B.R. Bloem, Dr. J. Nonnekes, and Dr. J.P.P. Van Vugt. Research protocol - Vibrating socks for Parkinson's Disease. Technical report, Medisch Spectrum Twente, Radboudumc, Enschede, 2019.

[41] S. Janssen, J. J.A. Heijs, M. Bittner, E. Droog, B. R. Bloem, R. J.A. Van Wezel, and T. Heida. Visual cues added to a virtual environment paradigm do not improve motor arrests in Parkinson's disease. *Journal of Neural Engineering*, (4):046009, mar . ISSN 1741-2552. doi: 10.1088/1741-2552/ABE356.

[42] Sabine Janssen, Benjamin Bolte, Jorik Nonnekes, Marian Bittner, Bastiaan R. Bloem, Tjitske Heida, Yan Zhao, and Richard J. A. van Wezel. Usability of Three-dimensional Augmented Visual Cues Delivered by Smart Glasses on (Freezing of) Gait in Parkinson's Disease. *Frontiers in Neurology*, 8(JUN):1, jun 2017. doi: 10.3389/FNEUR.2017.00279.

[43] Sabine Janssen, Jaap de Ruyter van Steveninck, Hizirwan S. Salim, Helena M. Cockx, Bastiaan R. Bloem, Tjitske Heida, and Richard J.A. van Wezel. The Effects of Augmented Reality Visual Cues on Turning in Place in Parkinson's Disease Patients With Freezing of Gait. *Frontiers in Neurology*, 11:185, mar 2020. ISSN 16642295. doi: 10.3389/FNEUR.2020.00185/FULL.

[44] Monique Paulich, Martin Schepers, Nina Rudigkeit, and Giovanni Bellusci. Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications. Technical report, Xsens Technologies, Enschede.

[45] Mihaela I. Chidean, Óscar Barquero-Pérez, Rebeca Goya-Esteban, Alberto Sánchez Sixto, Blanca de la Cruz Torres, Jose Naranjo Orellana, Elena Sarabia Cachadiña, and Antonio J. Caamaño. Full Band Spectra Analysis of Gait Acceleration Signals for Peripheral Arterial Disease Patients. *Frontiers in Physiology*, 9(AUG):1061, aug 2018. ISSN 1664042X. doi: 10.3389/FPHYS.2018.01061.

[46] Julià Camps, Albert Samà, Mario Martín, Daniel Rodríguez-Martín, Carlos Pérez-López, Joan M. Moreno Arostegui, Joan Cabestany, Andreu Català, Sheila Alcaine, Berta Mestre, Anna Prats, Maria C. Crespo-Maraver, Timothy J. Counihan, Patrick Browne, Leo R. Quinlan, Gearóid Laighin, Dean Sweeney, Hadas Lewy, Gabriel Vainstein, Alberto Costa, Roberta Annicchiarico, Àngels Bayés, and Alejandro Rodríguez-Molinero. Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge-Based Systems*, 139:119–131, jan 2018. ISSN 0950-7051. doi: 10.1016/J.KNOSYS.2017.10.017.

[47] Daniel Godoy. Understanding binary cross-entropy / log loss: a visual explanation. URL `https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a`.

[48] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Annh Ratanamahatana, and Eamonn Keogh. The UCR Time Series Archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, oct 2018. ISSN 23299274. doi: 10.48550/arxiv.1810.07758.

[49] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019*, nov 2017. doi: 10.48550/arxiv.1711.05101.

[50] Chetna Khanna. Independent and Identically Distributed, 2021. URL `https://towardsdatascience.com/independent-and-identically-distributed-ce250ad1bfa8`.

[51] Yi Xia, Jun Zhang, Qiang Ye, Nan Cheng, Yixiang Lu, and Dexiang Zhang. Evaluation of deep convolutional neural networks for detection of freezing of gait in Parkinson's disease patients. *Biomedical Signal Processing and Control*, 46:221–230, sep 2018. ISSN 1746-8094. doi: 10.1016/J.BSPC.2018.07.015.

[52] Thomas Bikias, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, and Leontios J. Hadjileontiadis. DeepFoG: An IMU-Based Detection of Freezing of Gait Episodes in Parkinson's Disease Patients via Deep Learning. *Frontiers in Robotics and AI*, 8, may 2021. ISSN 22969144. doi: 10.3389/FROBT.2021.537384.

[53] Alexandra-Maria Tautan, Alexandra-Georgiana Andrei, and Bogdan Ionescu. Freezing of Gait Detection for Parkinson's Disease Patients using Accelerometer Data: Case Study. pages 1–4, dec 2020. doi: 10.1109/EHB50910.2020.9280223.

[54] Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE*, 16(7):e0254841, jul 2021. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0254841.

A Results Wafer dataset

The Wafer dataset contains measurements recorded by one sensor during the processing of a silicon wafer for semiconductor fabrication. The two classes are normal and abnormal. The training set contains 1000 samples, the test set 6164 samples. The length of the samples is 152, the number of channels is 1. The data is imbalanced, 10.7% of the training data and 12.1% of the test data is abnormal [48]. Therefore, a class weight of 0.10 for the data of class 0 and a class weight of 0.90 for the data of class 1 is used in the CNN and InceptionTime.

Convolutional neural network
The CNN is trained in 100 epochs. Since the CNN is designed in this study, the results cannot be compared with another study. However, the results are comparable with other classification networks [20], and the performance is high with an accuracy of 99.8% and the ROC curve in figure 14.

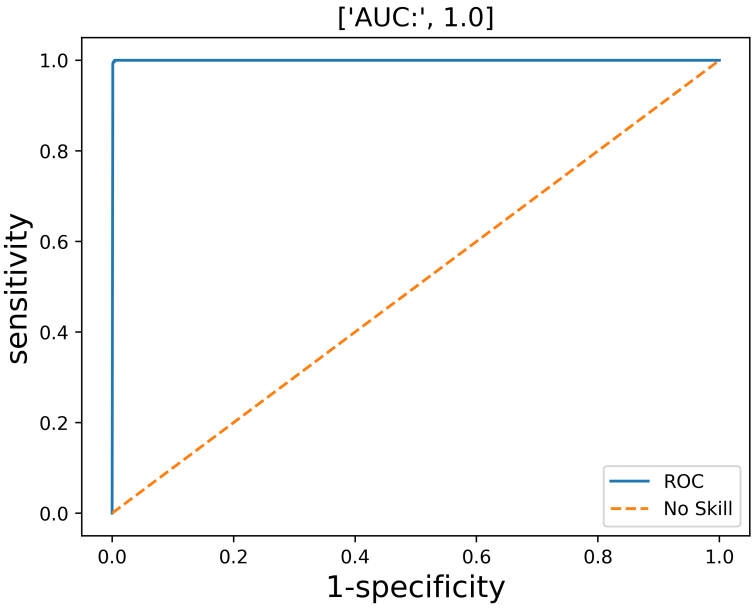


Figure 14: ROC curve of the CNN trained on the Wafer dataset.

MiniRocket
The ROC curve of MiniRocket trained on Wafer can be found in figure 15. The implementation of MiniRocket reached an accuracy of 99.0%, a little bit less than the 99.6% reported by Dempster et al. [18]. These differences may be due to optimization of hyperparameters.

InceptionTime
InceptionTime is trained in 100 epochs. The ROC curve of InceptionTime trained on the Wafer dataset is depicted in figure 16. An accuracy of 99.6% was reached, while Ismail-Fawaz et al. reported an accuracy of 99.9% [19]. This difference in performance may also be declared with optimization.

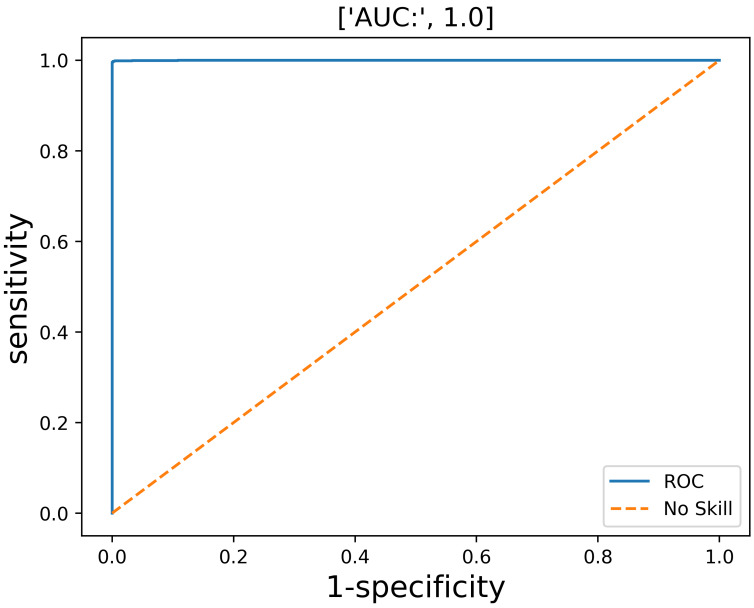


Figure 15: ROC curve of MiniRocket trained on the Wafer dataset.

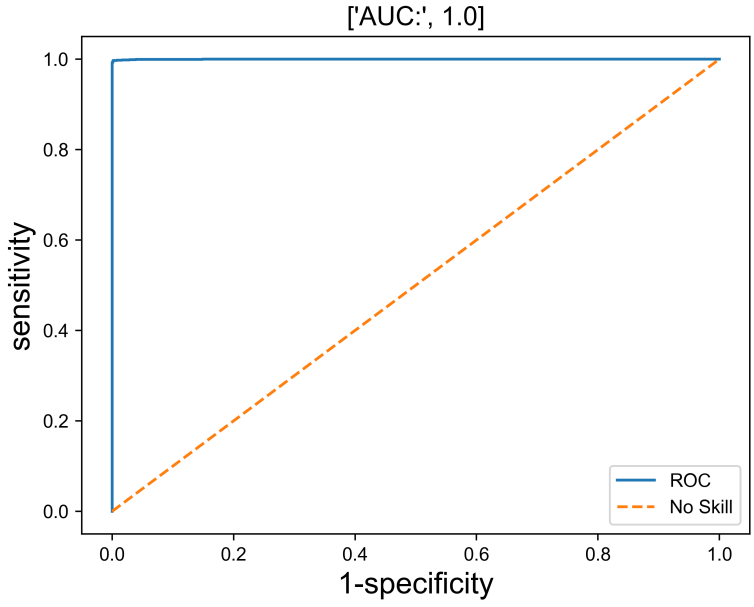


Figure 16: ROC curve of InceptionTime trained on the Wafer dataset.