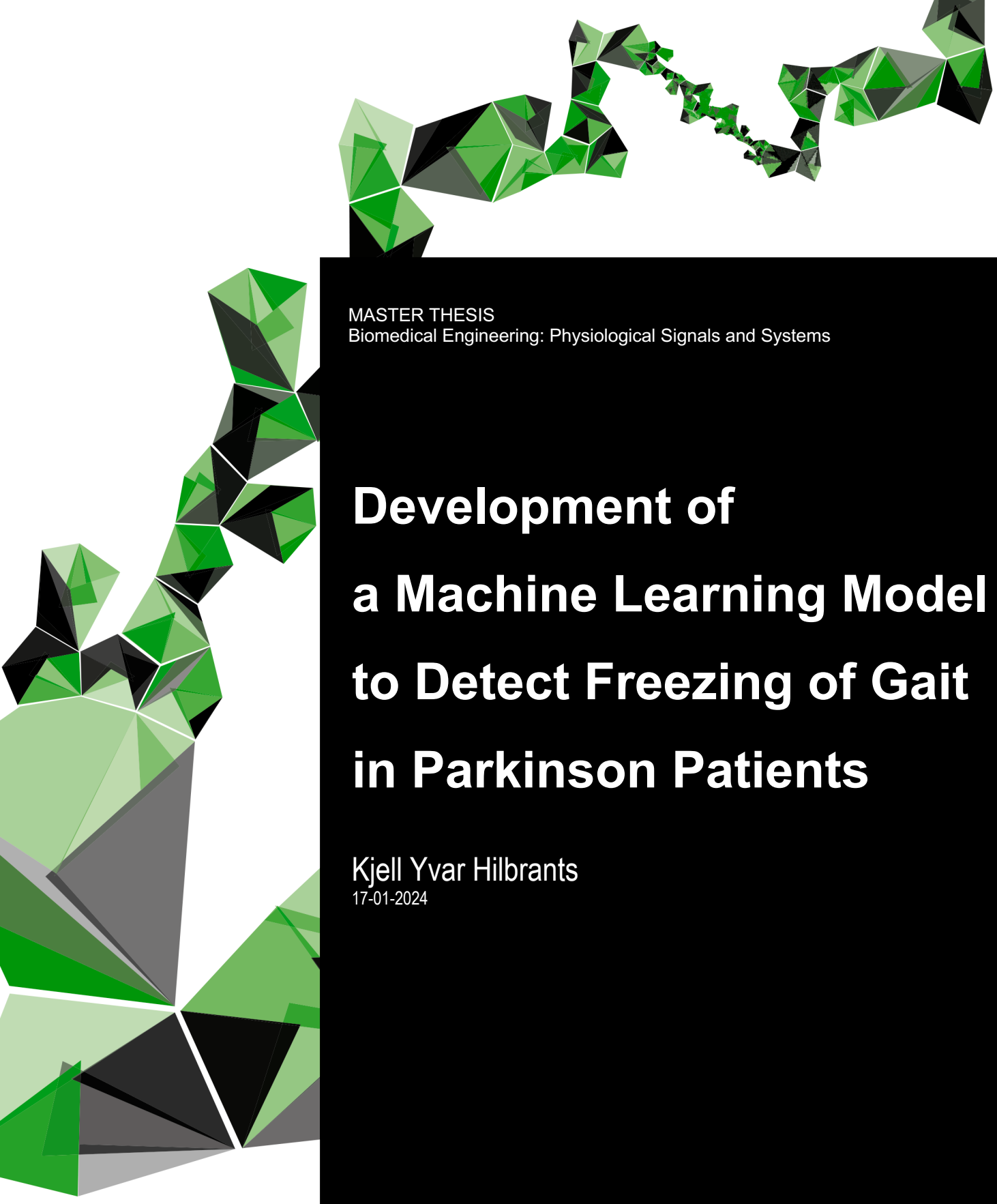


MASTER THESIS
Biomedical Engineering: Physiological Signals and Systems

Development of a Machine Learning Model to Detect Freezing of Gait in Parkinson Patients

Kjell Yvar Hilbrants
17-01-2024

TNW – Faculty of Science and Technology
Biomedical Signals and Systems (BSS)

Examination Committee

Dr. Ir. Ciska Heida	Committee Chair
Ir. Juan D. Delgado Terán	Daily Supervisor
Mhairi K. MacLean, PhD.	External Member

UNIVERSITY OF TWENTE.

ACKNOWLEDGEMENTS

During my thesis I've had the pleasure of working with many individuals, without whom this thesis would not have been possible. Therefore, I would like to dedicate this first page to acknowledge and thank them.

Foremost, I would like to express my sincere gratitude to Juan for his continuous support of my research and this final step in my academic journey. Even though this was the last step of my master's, I've still learned a lot during this last year. Doing my assignment from home came with its challenges and was sometimes difficult, but our down-to-earth (bi-)weekly meetings and mail contact always helped me get the guidance I needed to push on. I could not have imagined a better mentor for my final thesis, and I am proud of the research we have done together.

Secondly, I am sincerely grateful to Ciska for taking time out of her busy schedule to guide both me and Juan during our research. As a thoughtful instructor, you always supported me through the hard subject-matter by asking the difficult questions, while also reeling me back in when my scope would become too broad. Without your knowledgeable guidance, I might still be getting buried in all the variables and possibilities to consider, and want to thank you for lighting the road ahead.

Thirdly, I would like to give my thanks to Rai, for giving an outside look on my project. It can be easy to stick to established trends in the research, taking certain made choices as set in stone without question. While we did not have many meetings, you were always enthusiastic to see my progress and asked interesting questions that hadn't come up before. I feel that, with those added insights, I was able to change my mindset and see relationships in the data that I hadn't before. Therefore, I'd like to thank you.

Finally, I would like to thank the entire PD research group for being the welcoming and supporting group they are. From my first day with you I felt seen and heard. It was heartening to see how, while everyone's individual research is somewhat different, we all heavily supported each other. A special thanks here goes to Emilie Klaver, for oftentimes going out of her way to help me with specific parts of my project. I give you all my heartfelt thanks, and wish you all the best!

LIST OF ABBREVIATIONS

5Fold-CV	5-Fold Cross Validation
AAS	Adjusted Auditory Stroop
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AR	Augmented Reality
AUROC	Area under the Curve of the Receiver Operating Characteristic
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Network
Conv1D	1D Convolutional Layer
CV	Cross Validation
DL	Deep Learning
EHH	eHealth House
FoG	Freezing of Gait
fs	Sample Frequency
IMU	Inertial Measurement Unit
LOS-CV	Leave-One-Subject-Out Cross Validation
MDS-UPDRS	Movement Disorders Society Unified Parkinson Disease Rating Scale
Mini BEST	Short form of the Balance Evaluation Systems Test
ML	Machine Learning
MLP	Multilayer Perceptron
NN	Neural Network
OFF-state	Unmedicated Dopaminergic OFF-state
ON-state	Medicated Dopaminergic ON-state
PD	Parkinson’s Disease
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SHE	Simulated Home Environment
STD	Standard Deviation

ABSTRACT

Background: Freezing of gait (FoG) is a highly incapacitating motor symptom of Parkinson's Disease (PD), affecting on average 50% of early and 80% of advanced PD patients. It is characterised by a brief episodic absence or marked reduction of forward progression of the feet despite the intention to walk. Episode manifestation and frequency depend heavily on situation, environment, and patient. Ambulatory cueing could serve as symptomatic treatment of FoG, but detection techniques are needed to facilitate this. Convolutional neural network (CNN) based architectures using inertial measurement unit (IMU) data have shown promise in solving this problem. Therefore, this research sought to develop such a classification model to detect FoG episodes using IMU data of the ankle.

Methods: Two seperate main datasets were utilised to individually train and test the developed architectures. One dataset consisted of measurement data acquired using a simulated home environment, while the other was made up of data from four previous studies focusing on gait tasks such as walking, turning, and navigating narrow pathways in a (clinical) lab setting. Data from 20 and 64 subjects was obtained for these datasets respectively, translating to 63.13 and 21.41 hours of data of which 2.70 and 2.52 hours contained FoG. A subdataset of the first main set was generated as well, consisting of only walking and FoG data, to also test model performance on a set consisting only of active movement data. Data was split into windows of 2 s with 75% overlap. After preprocessing, rotational data augmentation was used to balance classes for each dataset. Two architectures were developed, a lightweight Mono-Headed architecture and a more complex version in the Multi-Headed architecture. Model performance was evaluated using the area under the curve of the receiver operating characteristic (AUROC). Furthermore, both 5-Fold cross validation (5Fold-CV) and leave-one-subject-out cross validation (LOS-CV) were used to evaluate overall performance, generalisability, and individual patient adaptability.

Results: Both architectures performed similarly well on all tested datasets, with average AUROCs all >91.5% for 5Fold-CV and >88.9% for LOS-CV. When compared with a previous study, an improvement of 19% in 5Fold-CV AUROC is found on the same dataset. LOS-CV showed all obtained ROCs contained a densely packed cluster of subjects above the mean curve, with a more spread out minority of subjects under the mean curve. This indicates that developed models perform well for most subjects, while strugling with some. No sign of model instability was found, and two out of the three dataset trained models showed no to limited signs of overfitting.

Conclusion: Both developed architectures have high potential to be implemented in further research and show promise for use in a wearable device to facilitate on-demand monitoring and intervention of FoG in PD patients. By utilising only one sensor, which was rated highly wearable by PD patients, positive adaptation of the technology is deemed likely.

CONTENTS

1 INTRODUCTION

Parkinson's Disease (PD) is a complex and progressive degenerative condition of the brain associated with motor impairments and a wide variety of non-motor complications. The prevalence of PD increases with age, and has doubled globally in the last 25 years. Global estimates in 2019 indicated that 8.5 million individuals were affected by the disease [1]. Healthcare usage in recent years indicates trends in incidences ranging from 5 to more than 35 per 100,000 new PD cases yearly. With the overall aging average of the global population, PD prevalence will only increase dramatically in turn, expectantly doubling again in the next 20 years. [2]

1.1 Freezing of Gait

One of the incapacitating symptoms of PD is Freezing of Gait (FoG), characterised by a brief episodic absence or marked reduction of forward progression of the feet despite the intention to walk. This feeling is usually described by patients themselves as 'Feeling as if their feet are glued to the ground'. [3] FoG has been shown to affect about 50% of early and 80% of advanced PD patients. [4] The spectrum of when and how patients experience FoG is fairly broad, but FoG appearance can be divided into three general phenotypes regarding movement of the legs; shuffling, trembling, and complete akinesia. See Figure 1.1 below. [5]

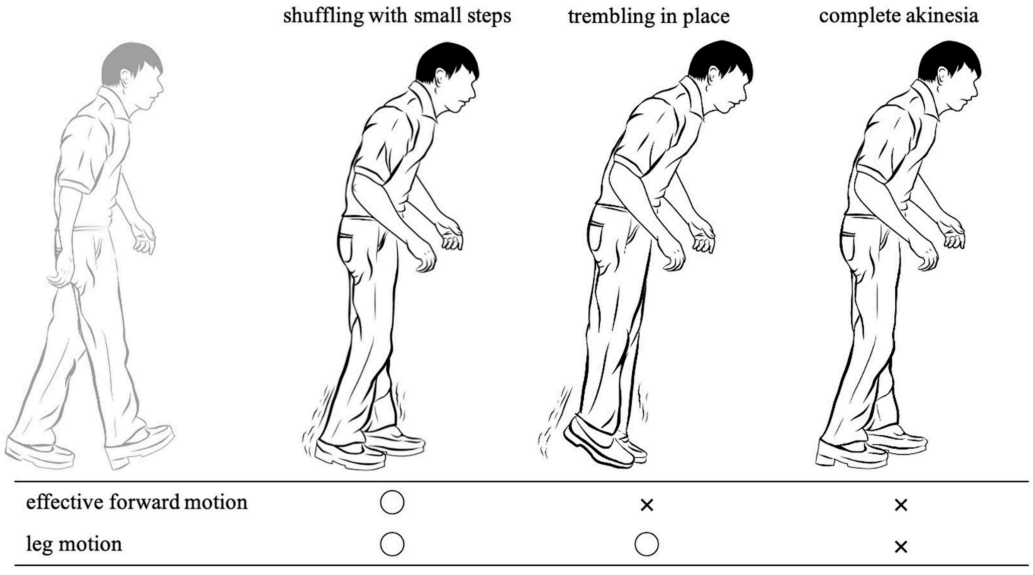


Figure 1.1: FoG divided into its three defined phenotypes; shuffling (maintaining effective forward and leg motion), trembling (losing effective forward motion, but maintaining leg motion), and akinesia (losing all motion). [5]

FoG manifestation and frequency seem to depend heavily on the situation. [6, 7, 8] Commonly known triggering motor actions are: gait initiation, turning, passing through narrow passages,

and approaching a destination (such as a chair). Furthermore, environmental, emotional, and cognitive factors also have known effects in triggering FoG. These include: approaching doorways or hallways, dual-tasking, distractions, anxiety, crowded places, and time pressure. Furthermore, characteristics can differ heavily between patients, some triggers will be less impacting for one than others. FoG can also be asymmetric, meaning that one leg is affected more than the other, leading to higher difficulty in, for example, turning to one side over the other. On average, FoG episodes will most commonly last only a few seconds or less, but are also known to be able to occasionally exceed 30 s in some patients. Markers for episodes can usually be observed before the actual onset of FoG. Gait can be seen to progressively deteriorate towards the start of an episode, cadence increases while the step length decreases. These effects on the gait are generally associated with the center of gravity falling forward over the feet. [8]
In affected PD patients, FoG is highly debilitating to their quality of life, activity levels, and further physical and mental health. [9] Patients have a heightened fall risk, effects of which are made even more severe regarding the already higher average age group of the PD population. [3, 10] This heightened fall risk, accompanied by fear of falling, are key factors in further lowering activity levels for many patients, leading to more sedentary lifestyles. Patients can become more dependant on others, and as a result often experience feelings of social isolation, anxiety and depression. [3, 11]

Currently, the first-line treatment of PD symptoms is pharmacotherapy with Levodopa. While on the medication (known as a patient's ON-state) significant decreases in both FoG frequency and severity have been shown, as well as improvement of other PD motor symptoms. However, the patient-specific nature of FoG leads to reduced to no effectiveness of the drug in some patients. [12]
Physiotherapy techniques, alongside medications, have also been shown to be helpful in further treating FoG. [13] One such useful technique is cueing via a myriad of stimuli. This can help focus the patient's attention on their gait, lowering or even preventing FoG and preserving functional gait. [13] Cues can take multiple forms, but they are all designed to stimulate the user via some external cognitive pathway, such as visual (e.g. lines on the floor indicating step length), auditory (e.g. metronome indicating cadence) or tactile (e.g. vibration indicating cadence). Besides relying on outside stimuli, patients can also actively cue themselves using techniques such as internal counting. These cues all serve to mirror an aspect of normal gait. This is believed to shift the user's habitual motor control to a more goal-oriented one, which can help with preventing and/or overcoming a FoG episode. [6] However, continuous offering of cues have shown diminishing returns in effectiveness, since the stimulus becomes monotonous and thus gets filtered out over time. Thus, ambulatory, or 'on demand', cueing could overcome this limitation by only offering cues when they are needed. Yet, to offer these ambulatory cues as a treatment, one would need to be able to determine when FoG episodes happen in an online setting. [14]

1.2 FoG Detection

To create a technology to facilitate ambulatory cueing, many studies have and are being done on FoG detection and/or prediction. Currently, the golden standard for studying FoG is via video annotation by experts. [15, 16] However, this practise is very time consuming, and furthermore would not be feasible in a prolonged daily living situation. Machine learning (ML) and Deep Learning (DL) techniques show significant promise in solving this problem. By utilising wearable sensors, ML and DL models can be trained to recognise markers for a FoG episode in the signal. [15, 16, 17]

A broad spectrum of biometric signals seem useful for this classification problem, such as heart rate, skin conductance, plantar pressure soles, and inertial measurement unit (IMU) data. [15, 18, 19] IMUs, capturing three-dimensional movement via accelerometer and gyroscope data, are most often used for automated detection of FoG. A recent review by Pardoel et al. [15] shows 60 out of their overall 74 reviewed studies to use either IMUs or one of their components (accelerometer or gyroscope) in both supervised and unsupervised settings. Besides prominence in the current scientific literature, IMUs are generally also highly wearable, relatively small, and can capture movement data of any body part.
As for ML and DL algorithms, Pardoel et al. [15] describe the highest performing FoG classifiers to be convolutional neural networks (CNNs), support vector machines, random forests and AdaBoosted decision trees. Out of all four of these algorithms, CNNs have the advantage of not needing any feature extraction because of its DL characteristics. CNNs are also by far the most widely popular, both within FoG detection and in other applications, because of their local pattern recognition strengths. [15, 20] However, as with all ML and especially DL techniques, large and balanced datasets are needed to turn developed algorithms into usable robust models. Both of these aspects seem lacking within the current FoG research environment, the largest dataset within the review containing 32 patients and FoG generally being undersampled because of its episodic nature and diminished frequency in lab settings [6]. Thus, efforts should be made to improve these two aspects when seeking to develop a FoG detection model. [15]

1.3 IMU Wearability

In 2022, O'Day et al. [16] sought to find the optimal harmony of IMU placement(s) and patient preference, since patient adherence to reliably wear the device(s) is critical to any future intervention's functioning. Using a relatively simple CNN, they achieved best results when using three IMUs, one in the lumbar region and one on each ankle. This three sensor leave-one-subject-out cross validated (LOS-CV) model attained an area under the receiver operating characteristic (AUROC) of 0.83. Furthermore, based on their wearability survey, this IMU configuration also utilised two of the three highest scoring sensor placements, and was the favourite subset among their sets consisting of three IMUs. Their best scoring minimal IMU set, consisting of the fewest amount of IMUs while still attaining LOS-CV AUROC performance within 5% of the previously mentioned best model, consisted of a single ankle sensor. While this placement was not the highest preferred among all single sensor placements, it did outperform the preferred single lumbar sensor. This minimal model attained an LOS-CV AUROC of 0.80, which is only 3.9% lower than the previously mentioned best set of three IMUs. [16]
Naturally, to least impact a user's quality of life, an intervention using only the minimal amount of sensors during day-to-day life would be preferred. Despite the ankle being ranked lower in wearability than the wrist for 1-IMU sets, overall individual wearability shows that these two were rated equal. This aspect, together with both high performance and minimal impact, suggests positive adaptation of an intervention using only a single ankle sensor for FoG in PD. [16]

1.4 Previous Work by Irene Heijink [21]

Previous to the current research, another master thesis on this subject was done by Irene Heijink [21]. Her aim was to determine how FoG could be detected using different pre-existant CNN based architectures using IMU data. Three classification models were analysed: a self-adapted CNN, MiniRocket [22], and InceptionTime [23]. The latter two are state-of-the-art time-series classification models chosen based on their performance when compared with other models tested on 26 multivariate time series datasets. [24] These three models were evaluated on

a dataset made up of lab measurements from four previous studies, which is also one of the datasets used in the current study. Furthermore, three different IMU sensor combinations were studied, with the aim of finding the combination with superior performance. Her findings showed the self-adapted CNN in combination with IMU data from lower legs and feet sensors performed best, attaining a 5-Fold cross validated (5Fold-CV) AUROC of 0.72. Yet, she also notes that learning behaviour often seems erratic or random, indicating that the model still has to guess at points. However, noting that this architecture performance shows potential, she thus called for further research to be done on the matter.
The current study aims to further build upon the findings of Heijink, taking inspiration from the architecture of her best performing self-adapted CNN and proposing the use of a data augmentation technique to partially solve the undesired learning behaviour. Furthermore, efforts are made to further minimise her proposed best IMU setup, only using one IMU instead of two, to maximally increase potential wearability. Proposed models of this study will be evaluated using multiple separate datasets, one of which being the same lab dataset used by Heijink, to directly compare findings with Heijink as well as with a completely separate dataset.

1.5 Research Aim

The aim of this research is to develop a ML model to detect FoG episodes in PD patients using a minimal IMU setup, in essence continuing parts of the work done by Irene Heijink in her master thesis [21]. Previous studies have shown CNNs to work well with this kind of local pattern recognition task both in PD research and other similar applications, thus this type of architecture will be used as a base focal point of the model. A single IMU at the right ankle will be used to supply movement data, since this placement has shown promising results on both performance and patient wearability rating. By using such a minimal setup, the way can be paved for further development into a highly sought after day-to-day wearable ambulatory cueing technology. Lastly, to mitigate influences from the difference in FoG manifestation characteristics, two separate datasets will be utilised to evaluate performance on both lab data and data from a simulated home environment. Thus, the research question is defined as:

How can a Convolutional Neural Network based Machine Learning Model be Developed to Detect Freezing of Gait in Parkinson Patients using Inertial Measurement Unit Movement Data from the Ankle?

2 BACKGROUND

This chapter provides needed background information and theory on technical aspects and initial choices made within this research.

2.1 Classification

The needed detection of FoG episodes described in this research is a classic binary classification problem, determining the presence or absence of FoG. ML and DL classifiers can generate probability distributions of these two classes from input data, showing which class the input data most likely belongs to. By supplying the classifier with a large robust dataset, hyperparameters can be adjusted over time to improve classification performance.

2.1.1 Machine Learning vs. Deep Learning

Information-rich markers in the input data can often be complex in nature, making them hard to discern for humans. However, an artificial intelligence (AI) is able to access and utilise these complex markers to, for example, classify time-series data as FoG or nonFoG in this research. ML is a subset of AI in which the algorithm has built-in capability to automatically learn and improve on its given task without human intervention. ML classification methods can utilise structured data, such as pre-extracted features, to output classes via a human-programmed prediction mechanism. An example of such an algorithm are decision trees, which are tree-like models where class probability choices are made using feature value thresholds at each branch, where branch order and thresholds are the learned parameters.
Going one step further, DL is a subset of ML which utilises neural networks (NNs) to essentially mimic human brain-like behaviour. These NNs contain a large amount of learnable parameters and hyperparameters to adapt to their given task and available data. Using this abundance of parameters they are able to translate complex markers into lower dimensional features, thus not requiring pre-extracted features from data. Thus, these algorithms do not have to be limited by predetermined lower dimensional feature choices and can instead utilise complex data directly. Yet, while both (sub)sets of classifiers require big datasets to learn on, DL classifiers require especially larger amounts of data. This due to their aforementioned complexity and high amounts of parameters. Thus, a choice between utilising techniques from either (sub)set requires an informed assessment of the complexity of both the use-case and available input data.

2.1.2 Dataset Balancing

Due to the episodic nature of FoG obtaining well-balanced datasets is often problematic. This class distribution can be further skewed by different and less frequent manifestation of FoG within lab settings, which is where the majority of current FoG research is done. [6, 15, 25] Training a model with highly imbalanced classes would most likely lead to inducing a bias to the

over-represented class, thus lowering classification accuracy for the under-represented class. Furthermore, when working with (mini-)batches, gradients will not be able to be estimated correctly when batches are only (or mostly) made up of one class. This leads to model weights not being able to update correctly, thus hampering learning. Because of this, a binary dataset such as the ones used in this research would preferably be as close to an even 50/50 split as possible.

Multiple ML techniques are thinkable to facilitate this. Popularly, downsampling of the over-represented or upsampling of the under-represented class should be chosen when possible as these generally do not directly impact model stability. Yet, both downsampling and upsampling methods are not without risks. Which individual windows should be left out or duplicated for example, the wrong choice here could lead to information-rich windows being left out or noise-heavy windows being duplicated. Randomly choosing these windows each epoch can solve this problem, but for relatively big datasets this can lead to noticeably slower processing times. To mitigate these downsides, data augmentation of the FoG windows can be used to upsample the undersampled FoG class. Data augmentation techniques artificially expand dataset size by creating modified copies of existing data. This modification can be done in a variety of ways, such as flipping, perturbation, colour augmentation, and rotation to name a few. Rotational data augmentation of IMU data for FoG research has been shown to have high performance while attaining low processing times. [25] Furthermore, this method can be used during pre-processing, and thus will not impact individual epoch times apart from added time associated with training more datapoints.

2.2 Convolutional Neural Networks (CNNs)

A CNN is a DL algorithm, while originally designed for image data they can also be used with time-series data. [26] This capability is further illustrated by CNNs being at the forefront of FoG detection in recent research. [15] The overall algorithm is inspired by the neurons in the human brain and especially the visual cortex. Individual neurons respond to stimuli from small regions, called the receptive field, which together overlap to cover the full visual area. In this manner, higher dimensional data can be translated to lower dimensional features. [26] These lower dimensional features can then be used in a classification algorithm. A schematic overview of this process can be seen in Figure 2.1. CNNs are made up of a collection of different components, a short summary is given below for each individual component.

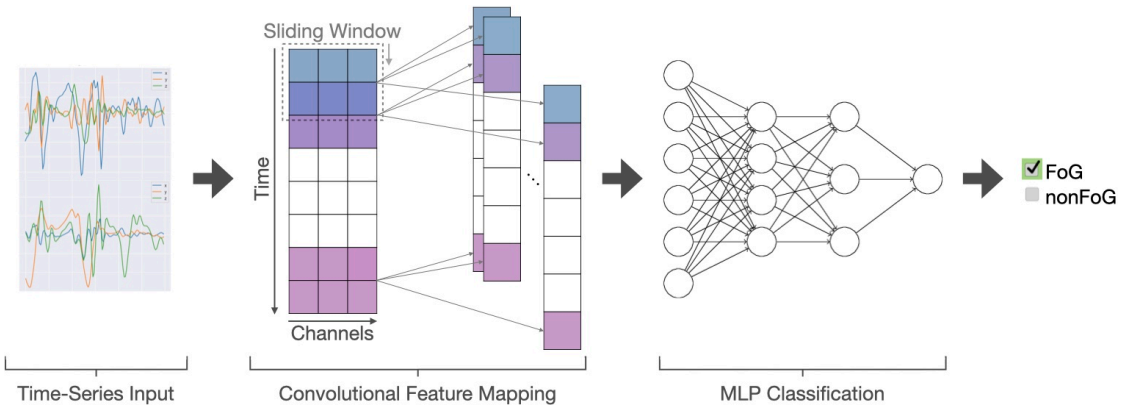


Figure 2.1: Schematic overview of a CNN classification algorithm used for this research. Multi-channel time-series input of the IMU is fed into the algorithm, this input is translated to a set amount of feature maps via convolution. Extracted features are then used to classify the input using a multilayer perceptron (MLP) to binary classes.

1D Convolutional Layer

The convolutional layer translates input data to a set of lower dimensional feature maps using convolution. A kernel size is set to determine the size of receptive field per convolution. This kernel is moved across the input data step-wise, computing one value on the feature map per kernel step, until all data has been seen at least once. A filter amount is set to determine how many feature maps will be produced. During training, filter weights are optimised to highlight the information-rich feature maps. [26]

ReLU Activation

The rectified linear unit (ReLU) activation function has become the default activation function for many different neural networks. The activation function outputs any positive value directly, while negative values are zero. By overcoming the vanishing gradient problem, models using this function are easier to train and often achieve higher performance. [27]

Batch Normalisation Layer

Batch normalisation stabilises the learning process by using mean and variance to normalise the data right before or after a non-linear function. By keeping the data distribution normalised to [-1,1], all segments of the model are able to learn faster since distributions will change less drastically due to parameter changes over each update cycle. Furthermore, model weights are kept small, promoting lower individual node dominance on decision making. [28]

1D Pooling Layer

A pooling layer further reduces data dimensionality after convolution. This decreases needed computational power and memory, and by consequence speeding up calculations. Furthermore, by reducing dimensionality, positional and rotational invariant dominant features are extracted, reducing noise. Pooling can either be done by taking the max or average of a portion of data. Max pooling in particular works as a noise suppressant, since it discards noisy activations altogether, while average pooling only suppresses by dimensionality reduction. For this added benefit, max pooling was chosen for this research. [26]

Dropout

Dropout is a regularisation method to reduce overfitting and increase generalisation. By randomly turning off a given percentage of units in the coupled layer, the model uses less nodes to train with, which in many cases increases generalisation. [29]

Dense Layer

Also known as the fully connected layer, a dense layer consists of n nodes, where each node is connected to all outputs from the previous layer. By stacking multiple dense layers a multi-layer perceptron (MLP) is formed. An MLP is a (usually) computationally cheap form of a classifier able to learn non-linear combinations of features in the output of the CNN. During a series of training epochs, the MLP can determine what combination of features holds the most information and sets its node weights accordingly using backpropagation. [26]

Sigmoid Activation

The last dense layer of the model uses a sigmoid activation function, which is specific to binary classification tasks. This activation function maps the output of the model to a probability of the input belonging to the first binary class ($P(Y = class1|X)$). A threshold for this probability can then be set to assign input measurements to either class. [30]

2.3 Adam Optimisation

Adam, derived from adaptive moment estimation, is a robust optimiser popularly used in ML and DL applications. The optimiser is an extended branch from the stochastic gradient descent (SGD) method. While SGD uses a fixed learning rate to update all weights during training, Adam makes use of adaptive learning rates. Calculation of these learning rates are based on the exponential moving average gradient and square gradients. The parameters $\beta 1$ and $\beta 2$ control the decay rates, thus gradient recollection, of these moving averages. Overall, Adam is a robust optimiser, computationally efficient, highly adaptable, and well suited for most problems. [31]

2.4 Binary Cross-Entropy (BCE) Loss Function

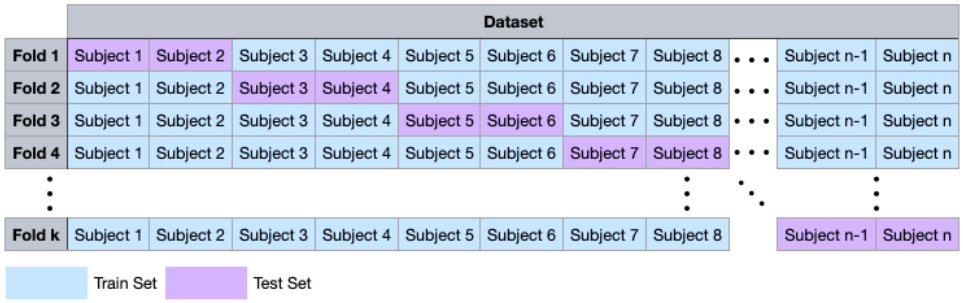
BCE (eq. 2.1), also known as log loss, is a loss function designed for binary classification problems. This loss function measures the discrepancy between ground truth labels (y) and generated probabilities (p), penalising outputted probabilities proportionally to distance from the ground truth. BCE is especially useful when prediction confidence is crucial, such as the medical diagnosis domain. Furthermore, BCE synergises well with sigmoid activation; probabilistic output can be used directly, the resulting gradient is conducive to learning, and wrong predictions with high confidence induce a high BCE loss due to their probabilities being close to 0 or 1. [32]

$$BCE(y, p) = -\sum_{i=1}^{n}[y_i * log(p_i) + (1 - y_i) * log(1 - p_i)] \tag{2.1}$$

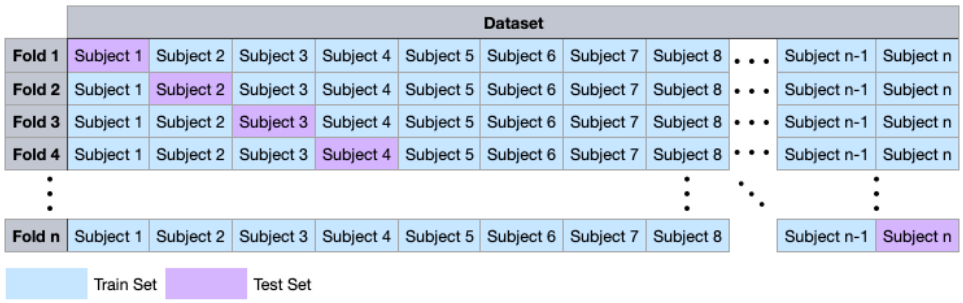
2.5 Cross Validation (CV)

CV is a very useful and commonly used statistical technique in ML for assessing performance of a trained model. [33] Various CV methods may be used, yet all have a similar algorithm at their core. First, the dataset is divided into two parts, one for training and one for testing. The model is trained on the training set. Then, the trained model is evaluated on the test set, which consists of not before seen datapoints. After performance metrics are obtained for that iteration, different parts of the whole dataset are chosen to be the new train and test sets and the process is repeated a number of times. Each iteration of this process is also commonly referred to as a 'Fold'. Obtained metrics from all folds can then be analysed on, for example, mean performance, variance, and stability. [34]

As mentioned, a variety of CV methods exist and are used for different use-cases. Within this research, two CV methods were selected fitting best with the use-case of FoG detection, namely 5Fold-CV and LOS-CV.
5Fold-CV is a form of kFold-CV with k=5, k=5 or 10 is usually preferred and taken as a general rule from empirical evidence. [33] This method divides the dataset into 5 subsets, each fold one of the 5 subsets is used as the test set and all others are used to form the train set.



(a) kFold Cross Validation



(b) Leave-One-Subject-Out Cross Validation

Figure 2.2: Schematic overviews of the two (a) kFold and (b) Leave-One-Subject-Out cross validation methods used in this research.

This process is then repeated until each subset has been the train set once. See Figure 2.2a. Generally, kFold-CV metrics are more stable and trustworthy than single iteration metrics, since bias is minimised by fitting and testing different parts of the dataset for each iteration. Thus, this method is representative of mean performance on the given population overall. [34]
LOS-CV is, in essence, a more specific type of kFold-CV, where k is equal to the number of subjects in the dataset. See Figure 2.2b. [34] As a result, the train set is maximised, ensuring as much data as possible is used to optimise model weights. Yet, simultaneously no data from the test subject is used for this optimisation, ensuring obtained metrics will be as unbiased to subject-specific markers as possible. [33] As a downside, LOS-CV can become quite computationally intensive for larger datasets, since each subject has a dedicated trained model on the large amount of data from all other subjects. [33] By lowering the amount of epochs, or using early stopping, runtime can be mitigated. Yet, this requires knowledge on required time for convergence of the model. Keeping this in mind, this CV method was chosen for model evaluation alongside 5Fold-CV considering FoG's patient-specific nature. It is thought that obtaining metrics per patient holds valuable information regarding generalisability and individual adaptability. Furthermore, obtained results from 5Fold-CV will be used to determine the amount of epochs for LOS-CV, mitigating unproductive runtime.

3 METHODS

For this research, two CNN based architectures were developed. Two separate datasets, consisting of 20 and 64 subjects, were utilised to train and test these models on. For each dataset, both models were cross validated in two ways to test overall performance, generalisability, and individual patient adaptability. Processing, development, and evaluation were all done using Python version 3.10 on a MacBook Pro (2018) using an Intel Iris Plus Graphics 655 1536 MB card and a Windows PC using an NVIDIA GTX 1080 card.

3.1 Data Acquisition and Datasets

Both datasets in this study were kept fully separate, training and testing an iteration of the model using subjects from only one set at a time. Within this study, these sets will be referred to as the simulated home environment (SHE) dataset and the Lab dataset. One of the main key differences between both datasets is the setting in which measurements were acquired. Measurements from the SHE dataset put the subject in a simulated home environment, whereas all measurements in the Lab dataset were done within a (clinical) lab setting. Since it is well-known within the research domain that FoG manifests differently and less frequent in lab settings [6, 15, 25], comparing similar models trained separately on either setting could give valuable insight into overall performance and applicability (i.e. usage for an intervention at home versus usage for FoG research within the lab). Further descriptions for both datasets can be found below.

3.1.1 SHE Dataset

The SHE Dataset consists of 20 self-reported subjects with PD who regularly experience FoG episodes daily. Subjects were invited to the eHealth House (EHH) in the Techmed Centre at the University of Twente, which is a lab setting which simulates a small home complete with kitchen, living room, bedroom, and bathroom. Each measurement day consisted of two separate parts, morning and afternoon, separated by a lunch break. Subjects were asked to skip their first Levodopa intake on the measurement day, thus facilitating measurements in dopaminergic OFF-state during the morning and ON-state during the afternoon. Subjects were fitted with multiple sensors, one of which being being the Movisens Move4 IMU at the right ankle used in this study. During both measurement parts, subjects underwent clinical assessments (e.g. MDS-UPDRS and Mini BEST), were asked to perform normal daily activities (such as their morning routine), and were accompanied on a walk around campus. Video of each measurement day was recorded in its entirety via cameras within the EHH and via a GoPro attached to the subject's chest aimed at their feet. Using these recordings, measurements were annotated offline by two experienced researchers reaching consensus and used to synchronise sensors.
The final SHE dataset consists of OFF- and ON-state measurements from 20 subjects. This translates to an overall 63.13 hours of available data, of which 2.70 hours are FoG episodes. In total, 898 individual FoG episodes were captured across all subjects of this dataset. For further analysis purposes, a copy of the full dataset has also been subdivided into a set containing only

instances of Walking and FoG. This set will be referred to as the SHE (Walking vs. FoG) set, and will serve to test performance when only taking active movement situations into account. This subset contains 10.91 hours of walking and 2.70 hours of FoG.

3.1.2 Lab Dataset [21]

The Lab Dataset consists of 64 subjects with PD and a recent history of disabling or regular FoG, and is combined from smaller datasets of four previous studies: Cinoptics [35], Hololens [36], Pedal [37], and Vibrating Socks [38]. All of these studies also included healthy controls, these were left out for this study. All measurements were done within controlled lab settings, such as empty hallways, and consisted of predetermined gait tasks such as walking straight forward, turning, navigating narrow passageways, and voluntary stopping. Furthermore, each study focused on measurements under different cueing conditions, for example tactile cueing for the Vibrating Socks study and augmented reality (AR) visual cueing in both Hololens and Cinoptics studies. Each study also contained control measurements for each subject where no cueing was present. IMU data was captured using the MVN Awinda Motion Capture System. For the current research, only the sensor at the lower right leg is used, since this placement is closest to that of the ankle sensor in the SHE dataset. All gait tasks were recorded on video and annotated offline by two independent raters per study. A short summary for each study is given below.

Cinoptics Study [35]

18 individuals with FoG were included from the Cinoptics study. Measurements were done at the end of the subject's dopaminergic medication cycle (end-of-dose). Three walking courses were completed in a 15 m long hallway, containing the following gait tasks: navigating a narrow passageway, stop-and-start, and turning. Furthermore, five different cueing conditions were tested throughout: two AR visual cues, one conventional visual cue, one conventional auditory cue, and no cue. Subjects were measured during two sessions, separated by a half hour break. Each session consisted of each cueing condition being tested subsequently using all three walking courses.

Hololens Study [36]

15 individuals with FoG were included from the Hololens study. Measurements were done at the end of the subject's dopaminergic medication cycle (end-of-dose). Subjects were asked to perform an 180° turn on the spot, under three different cueing conditions: one AR visual cue, one conventional auditory cue, and no cue. Fifteen turning trials were done for each cueing condition.

Pedal Study [21, 37]

7 individuals with FoG were included from the Pedal study. Measurements were done while subjects were in dopaminergic OFF-state after overnight withdrawal. Subjects were asked to walk down a straight 30 meter long hallway with two narrow passageways, then to make a wide turn at the end and return the same way. Each trial lasted approximately 30 seconds and multiple trials were performed back-to-back. Trails were performed under different cueing and cognitive conditions. To induce FoG, cognitive load was increased by use of the adjusted auditory stroop task (AAS). Each trial contained three to four randomly timed AAS tasks, where congruent word pairings (for example a male voice saying 'man') signaled the participant to start/continue walking and incongruent word pairings signaled the participant to stop.

Vibrating Socks Study [38]

24 individuals with FoG were included from the Vibrating Socks study. Measurements were done on two separate days, one for dopaminergic OFF-state and the other for dopaminergic ON-state. On each measurement day, while wearing a set of socks which can deliver vibrations in the subject's predetermined preferred cadence rhythm, subjects were asked to perform three gait tasks: walking 10 m, turning 360°, and a set gait trajectory. Four cueing conditions were tested: closed loop tactile cueing via vibrating socks, open loop tactile cueing via vibrating socks, auditory cueing, and no cueing.

Combining measurements from these four studies, the final Lab dataset consists of end-of-dose, OFF- and ON-state measurements from 64 subjects. This translates to an overall 21.41 hours of available data, of which 2.52 hours are FoG episodes. In total, 903 individual FoG episodes were captured across all subjects of this dataset.

3.2 Pre-processing

Both datasets were processed using almost identical protocols, with the Lab dataset having one extra step regarding artifact detection and removal. First, all data was synchronised with the corresponding video annotations. After this intial step, the aforementioned artifact detection and removal was done on the Lab dataset. For this, thresholds for acceleration and angular velocity were set to >100 m/s^2 and >1146 °/s (or >20 rads/s). Data spanning 5 samples before until 5 samples after the detected artifact was removed. These detection thresholds were already set and artifacts were removed from the dataset in a previous study, before acquiring the Lab dataset for the current study. [21] The SHE dataset was inspected as well regarding potential artifacts, but implementing a similar detection and removal scheme was decided unnecessary since included data did not seem to exceed given thresholds. Next, all data was filtered using a zero phase third order Butterworth band-pass filter between 0.3-15 Hz. This filter removed any potential drift and high frequency noise, while maintaining the full locomotor and FoG frequency bands (0.5-3 Hz and 3-8 Hz respectively) [39]. The filtered data was then resampled to 60 Hz, this served to normalise sample frequencies (fs) of both used sensors and maintained a high enough fs for movement analysis [17]. Lastly, inputs for the ML model were generated as 2 second long windows with 75% overlap and a step length of 0.5 s. This window length and overlap is very common within FoG detection research, and has generally been shown to achieve good results while maintaining low latency of detection. [15, 17, 40] Furthermore, by utilising a high overlap with a relatively small step length, more inputs will be created to train and test the proposed architectures with. [17] Windows were labeled FoG when $\geq$25% of samples (thus 0.5 s) contained FoG, this ensured that as many FoG windows as possible were generated which still contained enough information to train the model on. All inputs were saved per subject, to enable LOS-CV and kFold-CV where the model can be tested purely on data from previously unseen subjects. All end-of-dose, OFF- and ON-state measurements were used simultaneously, to ensure model performance would not be dependant on medicated state. [15] For the SHE dataset, morning (dopaminergic OFF-state) and afternoon (dopaminergic ON-state) were also linked, but per-window metadata annotations were made to enable analysis of model performance on either medicated state separately.

3.3 Data Augmentation

Because of the episodic nature of FoG, obtaining well-balanced datasets is often problematic. Class distributions of the SHE, SHE (walking vs. FoG), and Lab datasets are shown in Figure

3.1.

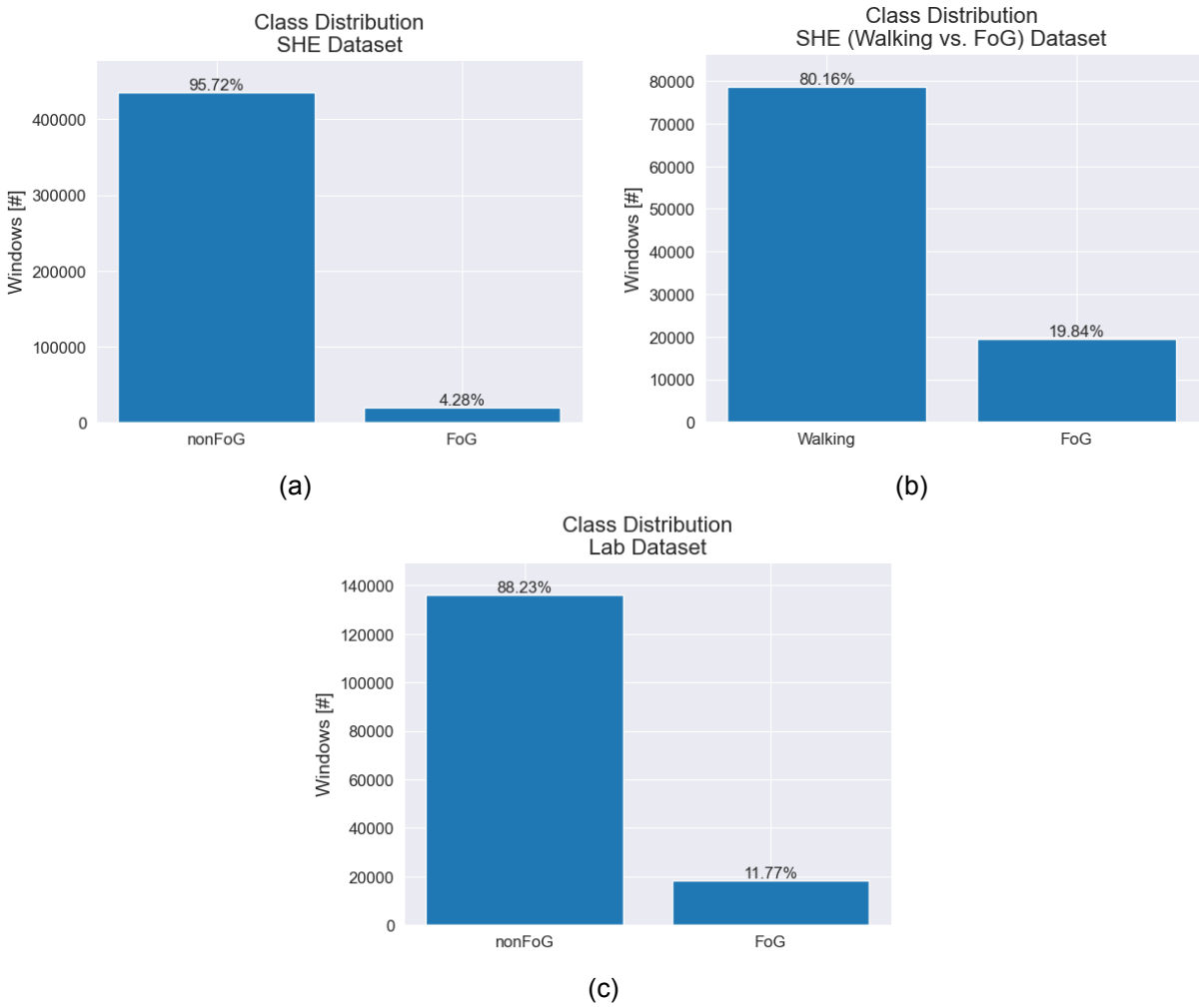


Figure 3.1: Class distributions of the pre-processed (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab datasets. All three sets show significant undersampling of the FoG class, meaning all three sets are highly imbalanced.

As can be seen, all sets are highly imbalanced, with the fraction of FoG windows only making up 4.28%, 19.84%, and 11.77% of each set. To balance these sets, a data augmentation algorithm by T. T. Um et al. [41] for augmenting IMU data was adapted to fit within this research paper's use-case. This algorithm performs a random 3D rotation within certain boundaries on the windowed IMU data to create a new synthetic data window, in essence simulating a rotation of the wearable sensor worn by the subject. Boundaries of the random 3D rotation were set to a full rotation in all three directions, this ensures that generated windows are, on average, as dissimilar as possible. Utilising this algorithm, the FoG class was upsampled, while the nonFoG class was left as is to ensure that potentially important information-rich windows were not discarded. Class distributions for each dataset were investigated regarding the amount of generated augmented windows per original FoG window, these FoG percentage curves can be seen in Figure 3.2.

To attain as close to an even split as possible, the constant for augmented windows per original FoG window was chosen accordingly. Thus, constants were chosen to be seven and three for the Lab and SHE (Walking vs. FoG) dataset respectively. The full SHE dataset, being the most imbalanced out of the three, would need 21 augmented windows per original to approach

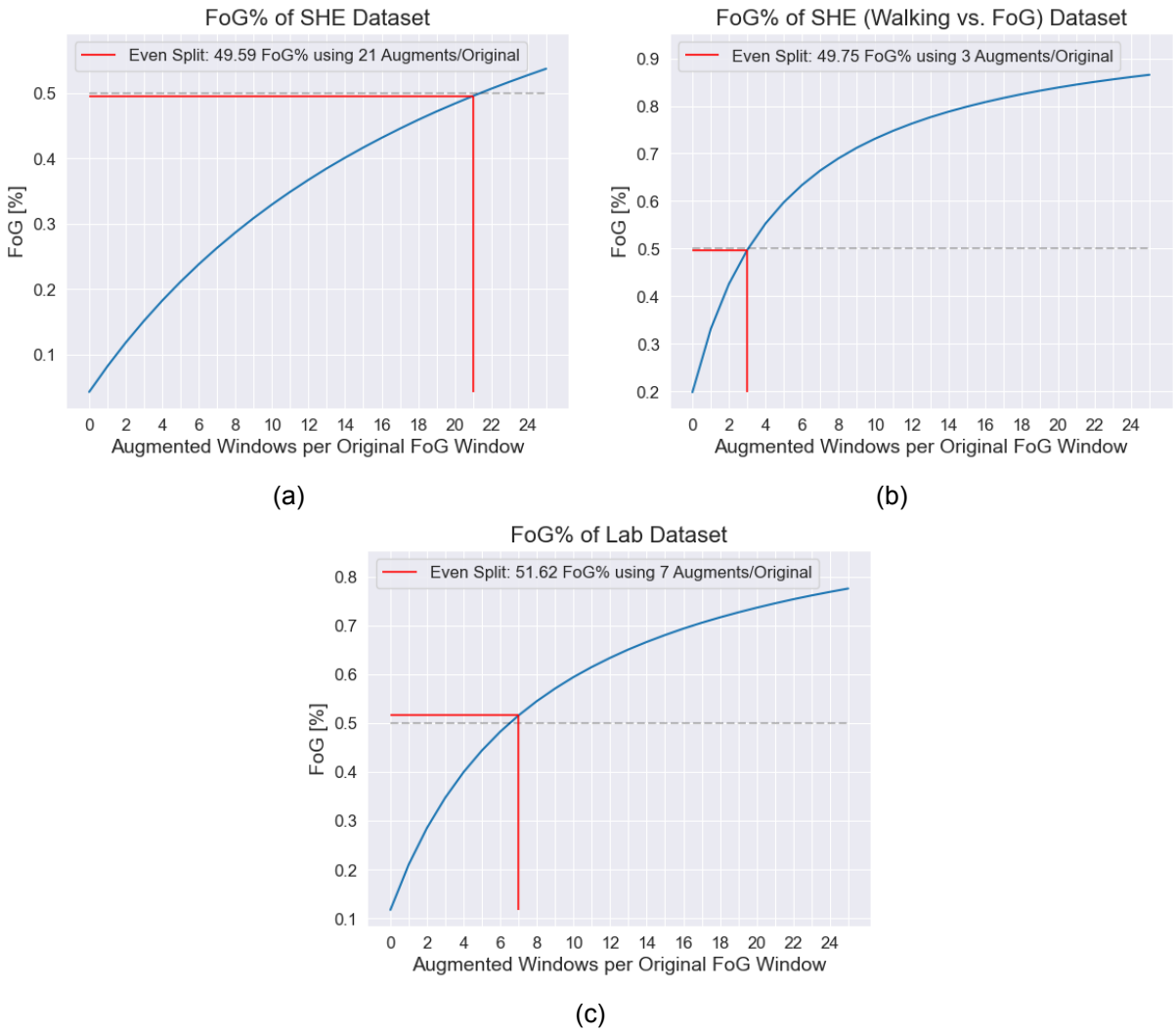


Figure 3.2: Investigated even split points of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab datasets. Indicating how many augmented windows per original FoG window would need to be generated to balance each set.

an even split. This is a much higher amount of augmented windows per original window than has been previously validated for the used augmentation technique. Therefore, as with the Lab dataset, the augmentation factor was set to seven for the full SHE dataset as well. This resulted in a close to 75/25 split for the SHE dataset. Resulting distributions of the augmented datasets can be seen in Figure 3.3.

3.4 Architectures

Two ML classification models were developed for this research. These models will be referred to as the Mono-Headed model and the Multi-Headed model, the Multi-Headed model being a more complex parallel version of the Mono-Headed model. As mentioned, inputs for these models consisted of 2 s long 60Hz time-series windows of pre-processed 6-channel IMU data. Both models were based on the use of a CNN to extract features from the data, which are then used to classify the window using a MLP. Adam optimisation was used with an initial learning rate of $5*10^{-5}$, and the loss function was set to binary cross entropy. Before each training cycle, data was shuffled to ensure a good distribution of both classes and batched into batches of 256

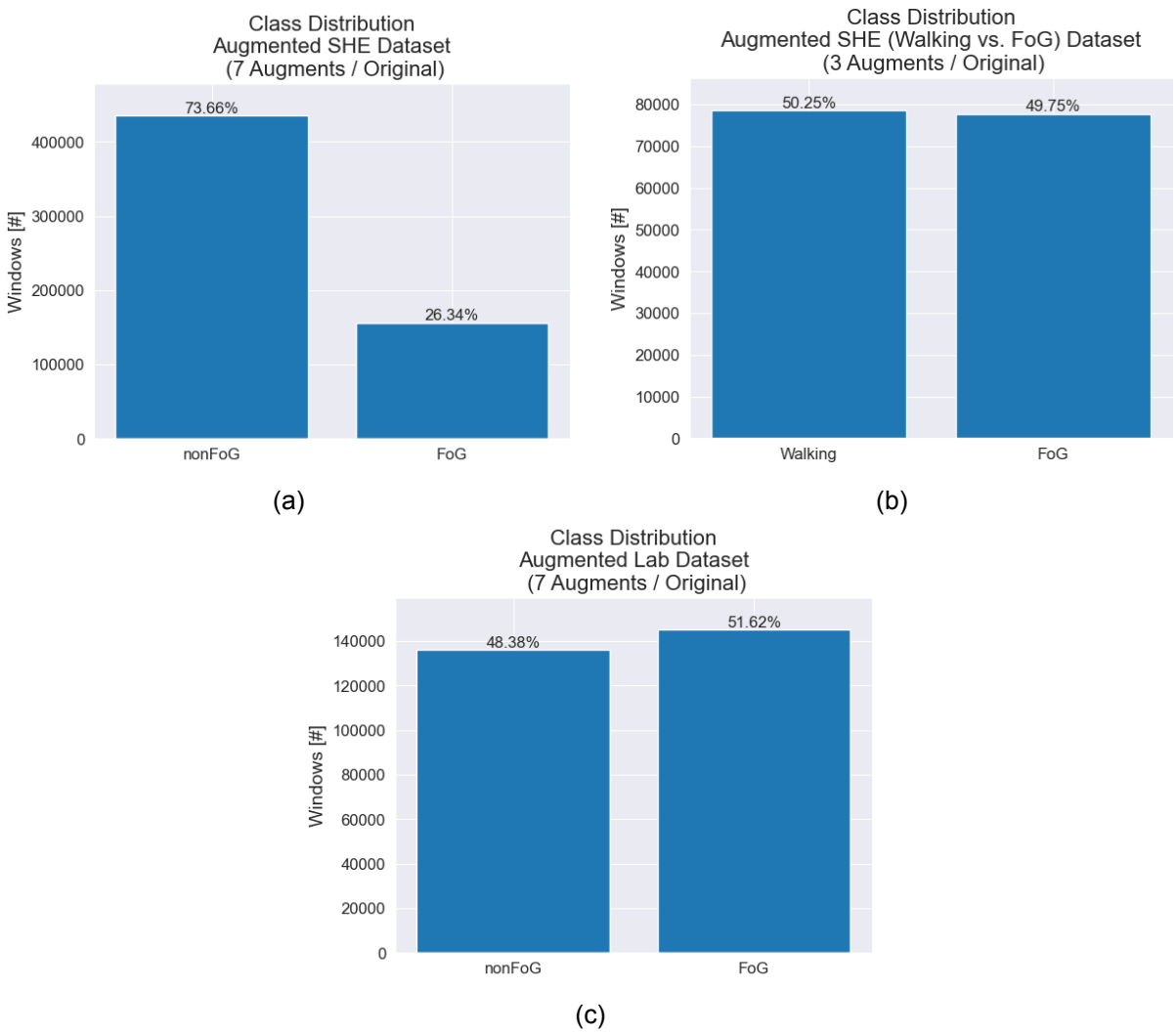


Figure 3.3: Class distributions of the augmented (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab datasets. Both (b) and (c) sets are well-balanced using 3 and 7 augmented windows per original FoG window respectively. Set (a) could not be evenly balanced without using too many augmented windows, thus 7 augmented windows per original were generated to achieve a close to 75/25 split.

inputs. Epoch length, either 100 or 30, was set depending on the used CV scheme. Models were developed, trained, and evaluated using TensorFlow for Python.

3.4.1 Mono-Headed

The Mono-Headed architecture was the first to be developed, with the aim to produce a model with high performance while also being computationally lightweight. A graphical overview of the architecture can be seen in Figure 3.4.
The first half of the model consists of three stacked CNN blocks. Each CNN block consists of a 1D convolutional layer (Conv1D) using ReLU activation, a batch normalisation layer, a max pooling layer, and dropout. Kernel size of the first CNN block is set at 15 (translating to 0.25 s), all other kernel sizes are set to 3. The amount of CNN filters doubles each block, starting at 16 filters for the first, doubling to 32 for the second, and doubling again to 64 for the third. Dropout increases each CNN block by 0.2, starting at 0.2 for the first, and increasing to 0.4 and 0.6 for the second and third. The amount of CNN blocks as well as all three named hyperparameters were tuned and optimised regarding both accuracy and AUROC performance during development.

Extracted features from the CNN are flattened and fed into the MLP half of the model. This MLP consists of three stacked fully-connected layers with decaying sizes. The first layer containing 64 nodes, the second 32, and the third 16. Each of these layers is also assigned a dropout of 0.5. Both size and dropout of the MLP were again optimised regarding accuracy and AUROC performance. Finally, binary outputs are returned using sigmoid activation.

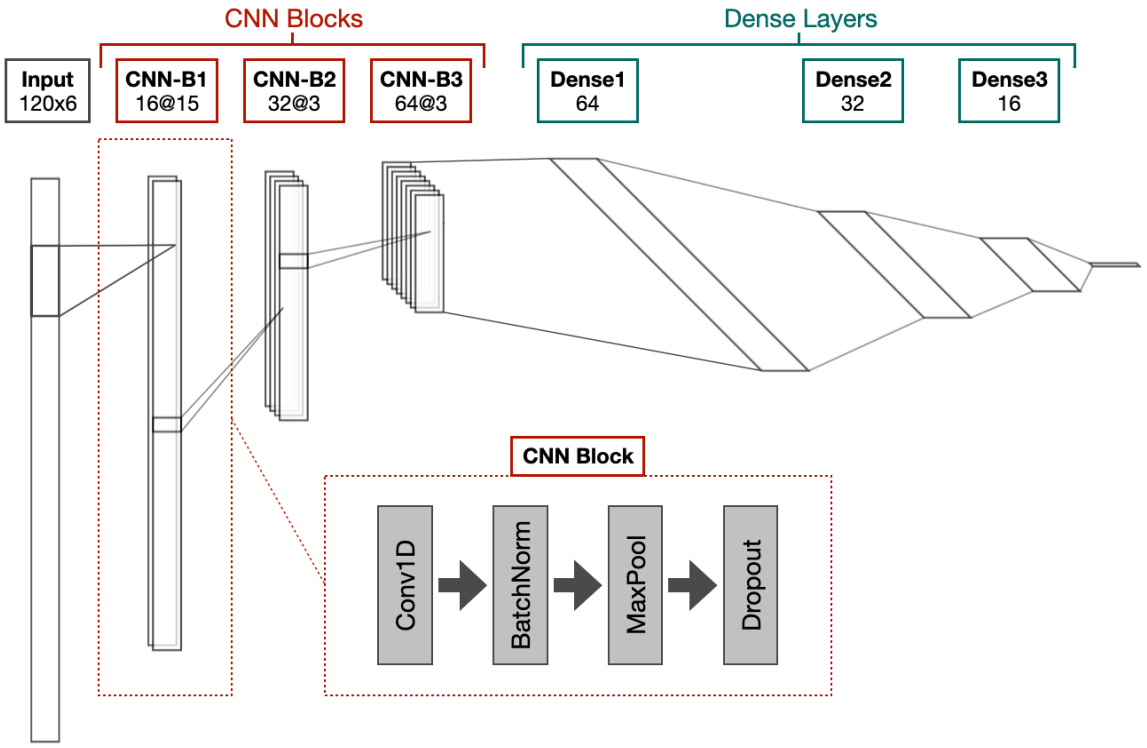


Figure 3.4: Schematic overview of the Mono-Headed architecture. Inputs of the model consist of time-series windows of 120 samples for 6 channels. The first half of the model consists of three stacked CNN blocks (CNN-B1, CNN-B2, and CNN-B3). Each block consists of a 1D convolutional layer (Conv1D), a batch normalisation layer, a max pooling layer, and a dropout layer. Filter amount and kernel size hyperparameters are given per block as Filters@Kernel. Output of the last CNN block CNN-B3 is flattened and fed into the second half of the model containing three dense layers. These dense layers decay in size, starting with 64 nodes in Dense1, then decreasing to 32 and 16 nodes for Dense2 and Dense 3 respectively.

3.4.2 Multi-Headed

The Multi-Headed model, being based on the Mono-Headed model, shares many similarities with the initial model. The main difference here is the three parallel CNN feature extractor heads instead of only one before the MLP. The number of heads, three, was chosen regarding promising results of a similar architecture proposed in a recent study in 2022 by Borzì et al. [29] The aim of this model was to test if a more complex model, which is able to simultaneously inspect data structures on three temporal resolutions, would improve classification performance. The drawback of course being higher computational load. A graphical overview of the architecture can be seen in Figure 3.5.
As mentioned, for this model the only differences are in the first half of the architecture. In essence, the CNN feature extractor of the Mono-Headed model is used in three parallel branches with each having its own copy of the input. The difference between these three branches being the kernel sizes of their first CNN blocks. These kernel sizes were chosen to be 7, 15, and 30

(translating to 0.125 s, 0.25 s and 0.5 s respectively). After flattening, the output of each branch is concatenated and fed into the MLP. Apart from this aspect, the overall structure is identical to the Mono-Headed model.

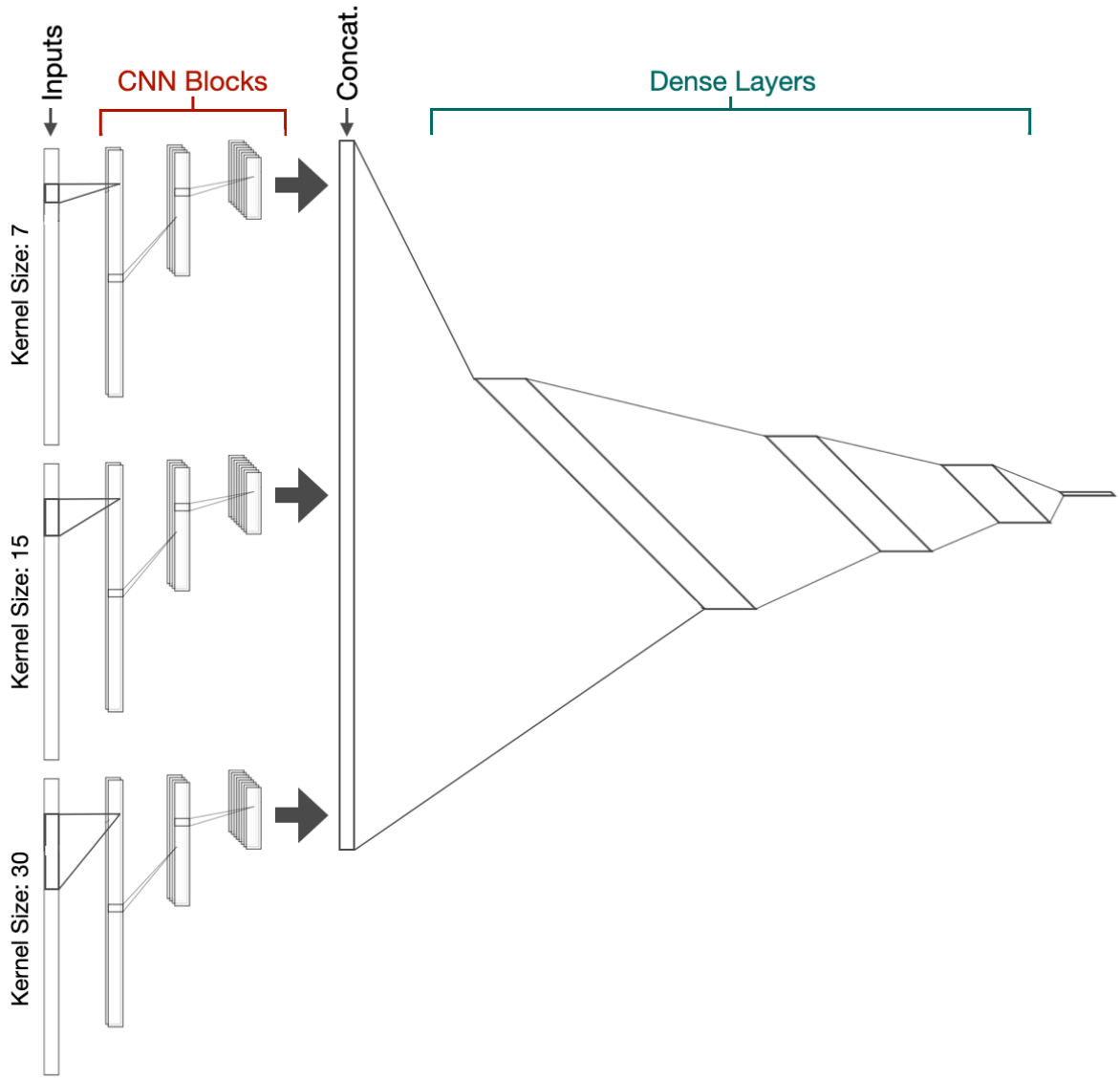


Figure 3.5: Schematic overview of the Multi-Headed architecture. Overall CNN and Dense halves are similar to those of the Mono-Headed architecture shown in Figure 3.4, i.e. each CNN branch in the first half consists of three CNN blocks of which the last output is fed into the Dense layers of the second half. Where the Mono-Headed architecture consisted of one branch of CNN blocks with an initial kernel size of 15, the Multi-Headed architecture consists of three separate branches with kernel sizes of 7, 15, and 30. Outputs of each branch are concatenated and then fed into the first Dense layer.

3.5 Validation and Analysis

All test results were evaluated according to the following metrics: accuracy (eq. 3.1), sensitivity (eq. 3.2), specificity (eq. 3.3), F1-score (eq. 3.4), and AUROC. The latter of which, AUROC, will serve as the main metric to evaluate overall model performance. AUROC in essence evaluates how efficient a model is in distinguishing classes, the closer to 1.0 meaning a more efficient model. Furthermore, AUROC considers possible trade-offs between both sensitivity and specificity, while other metrics like accuracy only show how many predictions are correct at a

predetermined class cut-off threshold. Yet, other metrics should also be given as context, since AUROC can be skewed by, for example, high dataset imbalance.

$$Accuracy = \frac{TP + TN}{P + N} \tag{3.1}$$

$$Sensitivity = \frac{TP}{P} \tag{3.2}$$

$$Specificity = \frac{TN}{N} \tag{3.3}$$

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.4}$$

Two CV schemes were implemented to test overall performance, generalisability, and individual patient adaptability. Firstly, using 5Fold-CV, each dataset was divided into 5 subject groups, where each group served as part of the test set once. 5Fold-CV models were set to train for 100 epochs, to ensure full model convergence and facilitate analysis of model learning behaviour and stability over time. Secondly, using LOS-CV, models were trained on data from all subjects apart from one, then tested on said left out subject. LOS-CV models were set to train for 30 epochs, this lower amount was observed to be a fitting epoch length for model convergence from the 5Fold-CV tests, and was thus chosen to lower overall computational load and runtime. For both CV schemes, it was ensured that data from a subject would only be present in either the train or test set per iteration, to not bias the test set.

4 RESULTS

Both the Mono-Headed and Multi-Headed architectures were tested extensively using all three sets separately. All subjects from both SHE and Lab datasets were used during their respective tests, 20 subjects for the SHE sets and 64 for the Lab set. Architectures were cross validated for both usage on subject groups (5Fold-CV) and per subject performance (LOS-CV), for which architectures were trained for 100 and 30 epochs respectively per fold. To enable calculation of LOS-CV average AUROC, F1-score and sensitivity, patients without FoG were disregarded for these metrics. The Mono-Headed architecture was relatively quick to train, with average epoch times being 3 minutes for the SHE set, 20 seconds for the SHE (Walking vs. FoG) set, and 35 seconds for the Lab set. On average, the Multi-Headed architecture was, as expected, much slower, taking about 2.5 times longer per epoch.

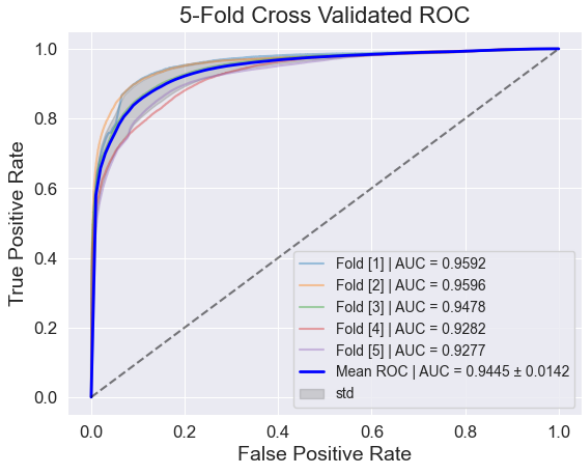
4.1 Mono-Headed

The Mono-Headed architecture attained average 5Fold-CV AUROCs of 0.9445 (STD±0.0142), 0.9564 (STD±0.0225), and 0.9209 (STD±0.0304) for the SHE, SHE (Walking vs. FoG), and Lab sets respectively. Furthermore, the obtained 5Fold-CV ROCs show all curves to be grouped close together, where the SHE set is most similar across all folds and the other two sets having some outliers. All obtained 5Fold-CV metrics for the Mono-Headed architecture can be found in Table 4.1.

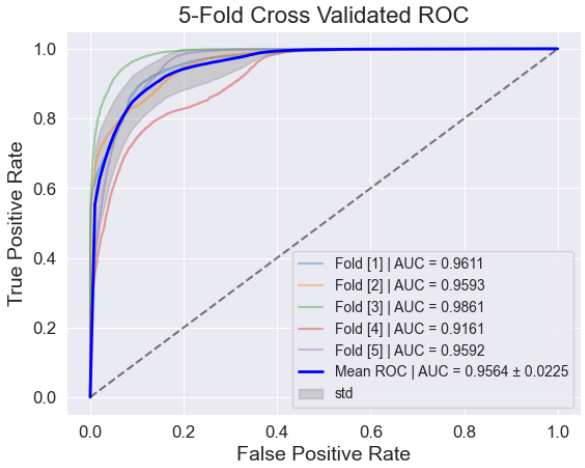
LOS-CV of the same architecture attained average AUROCs of 0.9145 (STD±0.0867), 0.8896 (STD±0.1463), and 0.9008 (STD±0.1139) for the SHE, SHE (Walking vs. FoG), and Lab sets respectively. Individual LOS-CV metrics per subject, as well as FoG distribution, can be found in Appendix A. As can be seen in Figure 4.2, individual subject ROCs are more spread than the obtained 5Fold-CV ROCs. It can be seen that, in all three graphs, there is a more dense cluster located in the upper left above the mean ROC and more spread out individual ROCs below. Furthermore, the obtained LOS-CV metrics in Table 4.2 similarly show larger STDs across all metrics.

Table 4.1: 5Fold-CV metrics of the Mono-Headed model

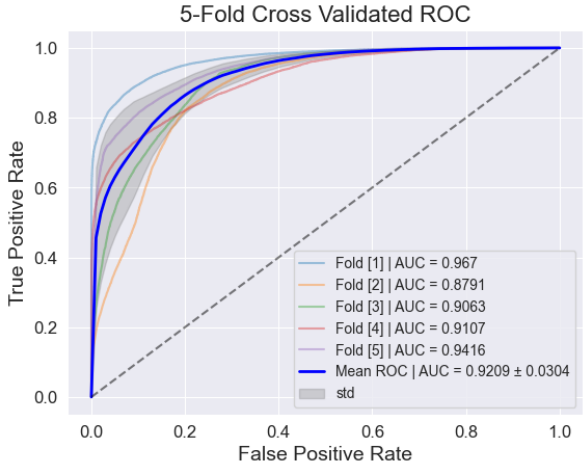
Set	Accuracy	Sensitivity	Specificity	F1-Score	AUROC
SHE	0.9002 (±0.0265)	0.7775 (±0.0761)	0.9432 (±0.0219)	0.7758 (±0.0428)	0.9445 (±0.0142)
SHE (Walking vs. FoG)	0.8764 (±0.0587)	0.9586 (±0.0225)	0.7964 (±0.0942)	0.855 (±0.0806)	0.9564 (±0.0225)
Lab	0.8254 (±0.0457)	0.9042 (±0.0379)	0.7479 (±0.0972)	0.8268 (±0.0773)	0.9209 (±0.0304)



(a) SHE



(b) SHE (Walking vs. FoG)

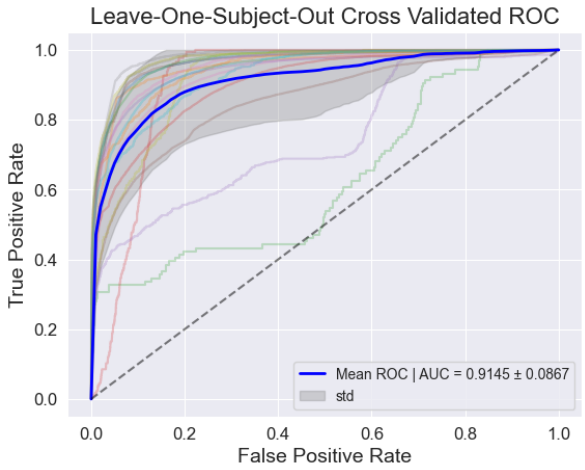


(c) Lab

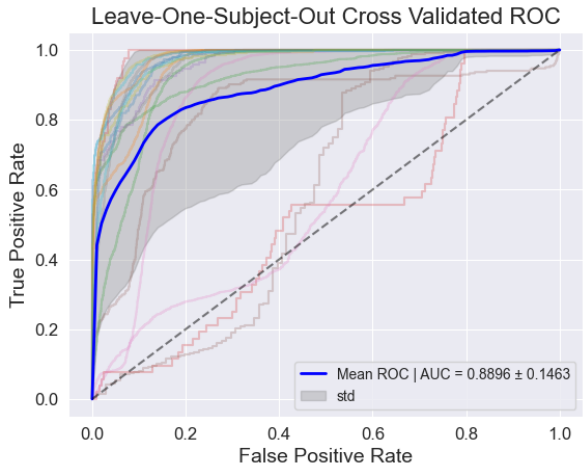
Figure 4.1: 5Fold-CV ROCs of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset trained Mono-Headed model.

Table 4.2: LOS-CV metrics of the Mono-Headed model

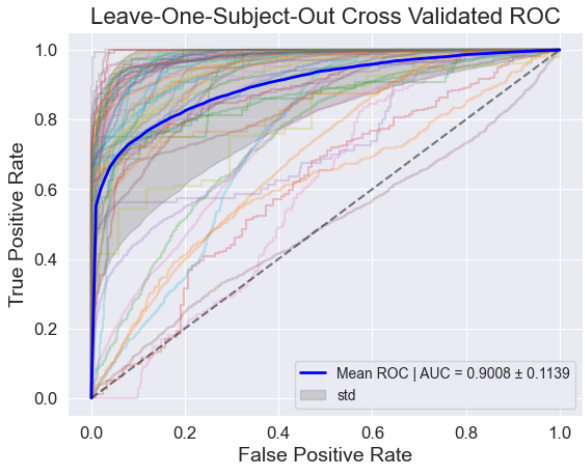
Set	Accuracy	Sensitivity	Specificity	F1-Score	AUROC
SHE	0.9109 (±0.0577)	0.6954 (±0.1922)	0.9387 (±0.0444)	0.6013 (±0.2951)	0.9145 (±0.0867)
SHE (Walking vs. FoG)	0.8598 (±0.1448)	0.9451 (±0.0745)	0.7372 (±0.2232)	0.6762 (±0.3217)	0.8896 (±0.1463)
Lab	0.8326 (±0.159)	0.9088 (±0.0761)	0.719 (±0.272)	0.7077 (±0.2768)	0.9008 (±0.1139)



(a) SHE



(b) SHE (Walking vs. FoG)



(c) Lab

Figure 4.2: LOS-CV ROCs of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset trained Mono-Headed model.

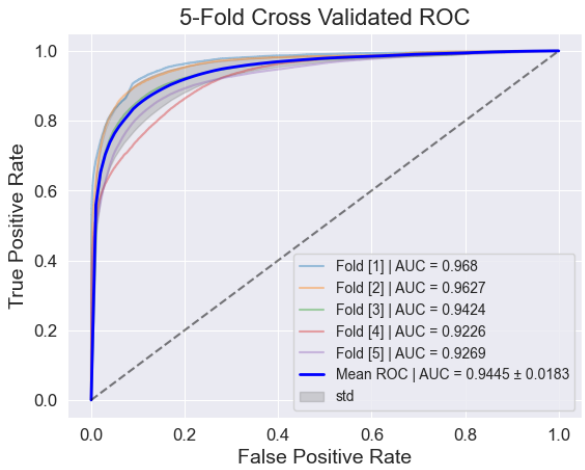
4.2 Multi-Headed

The Multi-Headed architecture attained average AUROCs of 0.9445 (STD±0.0183), 0.9491 (STD±0.0277), and 0.9157 (STD±0.0289) for the SHE, SHE (Walking vs. FoG), and Lab sets respectively. These average scores show on average similar performances when compared with the 5Fold-CV results of the Mono-Headed architecture. These similarities are further illus-

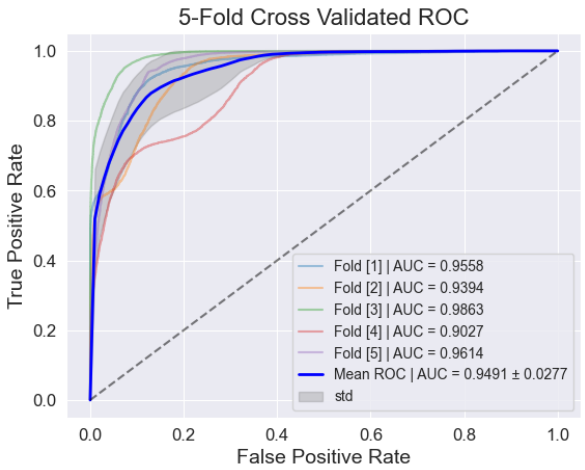
trated by the obtained 5Fold-CV ROCs shown in Figure 4.3, and all other metrics in Table 4.3.

Table 4.3: 5Fold-CV metrics of the Multi-Headed model

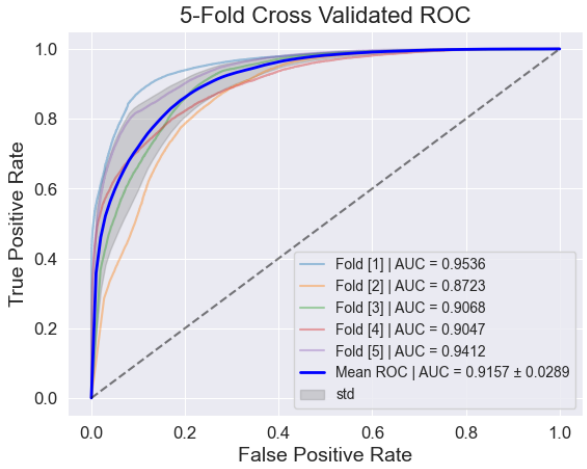
Set	Accuracy	Sensitivity	Specificity	F1-Score	AUROC
SHE	0.8983 (±0.0241)	0.791 (±0.0917)	0.9351 (±0.0263)	0.7718 (±0.0587)	0.9445 (±0.0183)
SHE (Walking vs. FoG)	0.8742 (±0.0597)	0.956 (±0.0276)	0.7878 (±0.1046)	0.8545 (±0.0776)	0.9491 (±0.0277)
Lab	0.8308 (±0.0465)	0.8937 (±0.0196)	0.7736 (±0.0659)	0.8292 (±0.0781)	0.9157 (±0.0289)



(a) SHE



(b) SHE (Walking vs. FoG)



(c) Lab

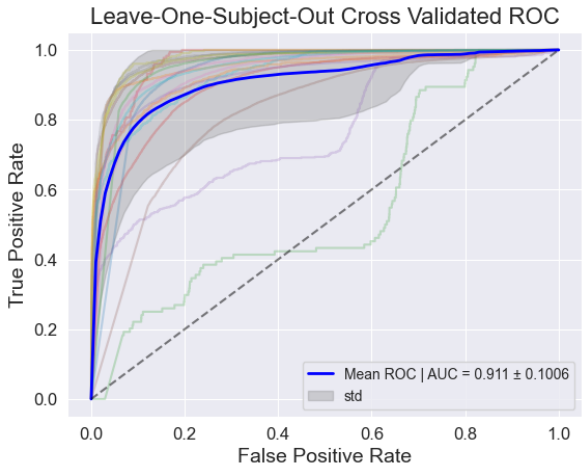
Figure 4.3: 5Fold-CV ROCs of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset trained Multi-Headed model.

In turn, LOS-CV of this more complex architecture attained average AUROCs of 0.911 (STD±0.1006), 0.8896 (STD±0.1496), and 0.9331 (STD±0.0656) for the SHE, SHE (Walking vs. FoG), and Lab set respectively. While both SHE sets attained similar AUROCs using LOS-CV on the Mono-Headed model, the Lab set shows noticeable improvement in both average and STD. This improvement can also be seen in Figure 4.4c, where less subject curves are present under the

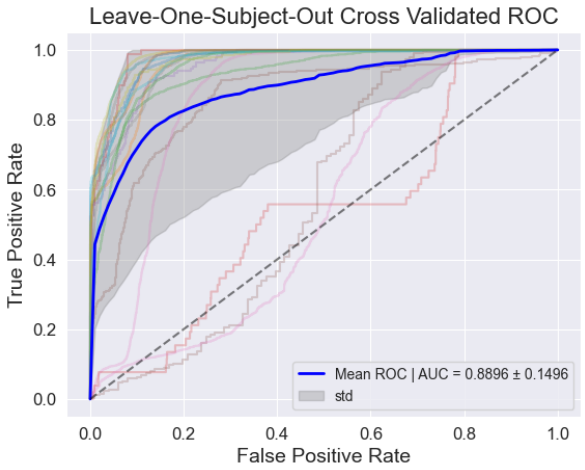
average ROC and STD area. This improvement can be seen in Table 4.4 as well, where average accuracy, and both average and STD of specificity all have better performance. Lastly, a higher STD for F1-score is also shown, this could be an effect of the higher specificities this architecture seems to reach for some subjects. All individual LOS-CV metrics per subject, as well as FoG distribution, can be found in Appendix B.

Table 4.4: LOS-CV metrics of the Multi-Headed model

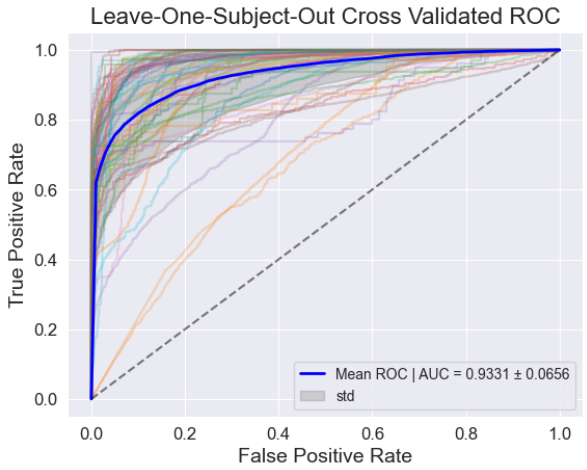
Set	Accuracy	Sensitivity	Specificity	F1-Score	AUROC
SHE	0.9004 (±0.0566)	0.7733 (±0.2227)	0.913 (±0.0604)	0.5981 (±0.297)	0.911 (±0.1006)
SHE (Walking vs. FoG)	0.8678 (±0.142)	0.9414 (±0.0826)	0.7453 (±0.2321)	0.6858 (±0.3176)	0.8896 (±0.1496)
Lab	0.8565 (±0.1446)	0.8945 (±0.0835)	0.7946 (±0.2173)	0.7403 (±0.2632)	0.9331 (±0.0656)



(a) SHE



(b) SHE (Walking vs. FoG)

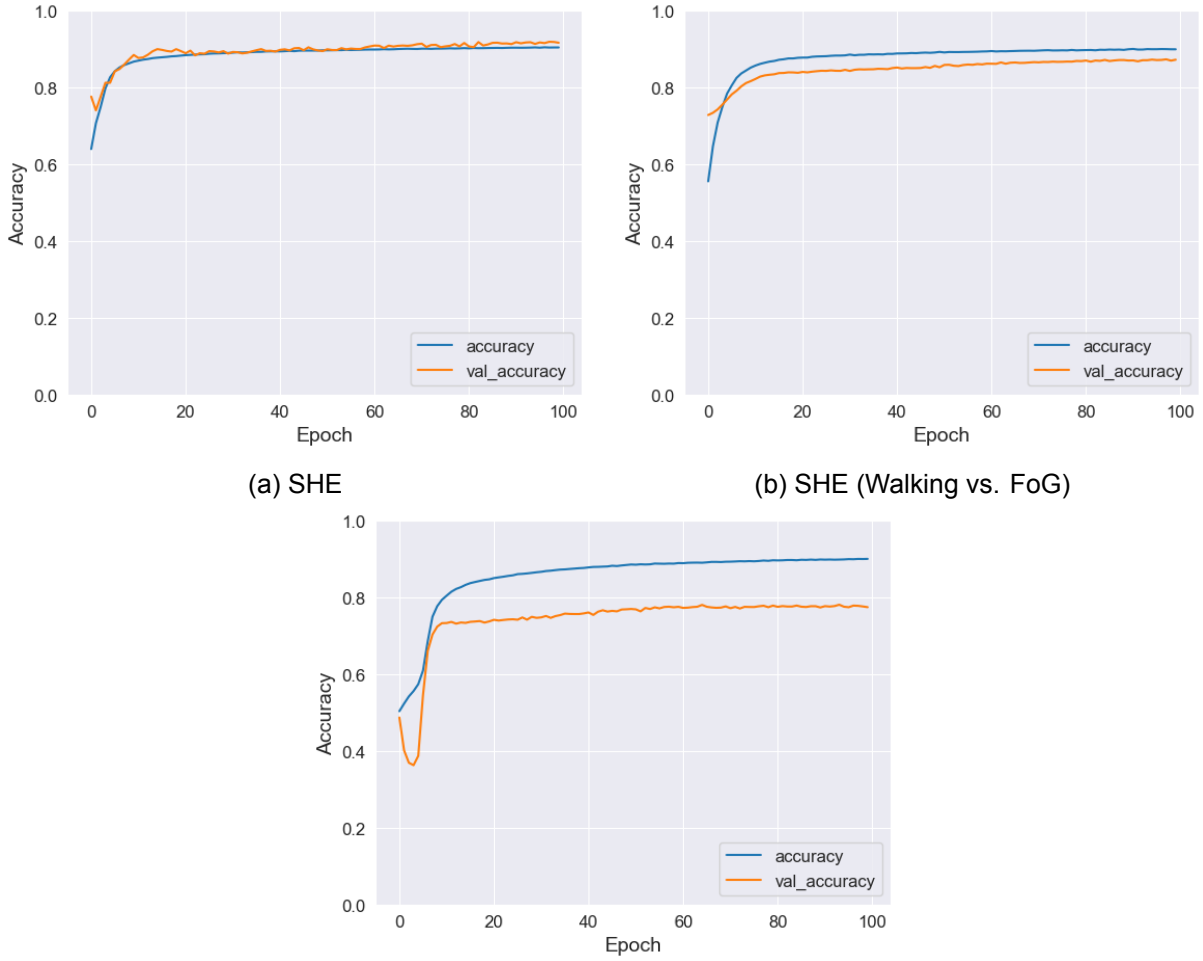


(c) Lab

Figure 4.4: LOS-CV ROCs of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset trained Multi-Headed model.

4.3 Overall Stability and Overfitting

Training and testing behaviour of both architectures showed similar signs regarding stability and overfitting. Both architectures show, on average, stable curves for training and testing accuracy and loss in all three sets. However, overfitting severity was noticeably different between the sets. Accuracy and loss learning curves of the Lab dataset models often showed signs of overfitting gaps between the train and test subsets. An example of this can be seen in Figures 4.5 and 4.6. For the SHE sets both accuracy and loss show (almost) overlapping curves for all epochs, with a slight gap for the SHE (Walking vs. FoG) set. This behaviour is very different in the Lab set, where big gaps can be seen in both accuracy and loss curves after approximately 7 epochs. Furthermore, loss drastically increases hereafter for the test set and stays high, while the training set loss continues to optimise. These signs of overfitting were generally seen across the majority of folds, thus showing that this is an inherent problem of the Lab dataset for these architectures.

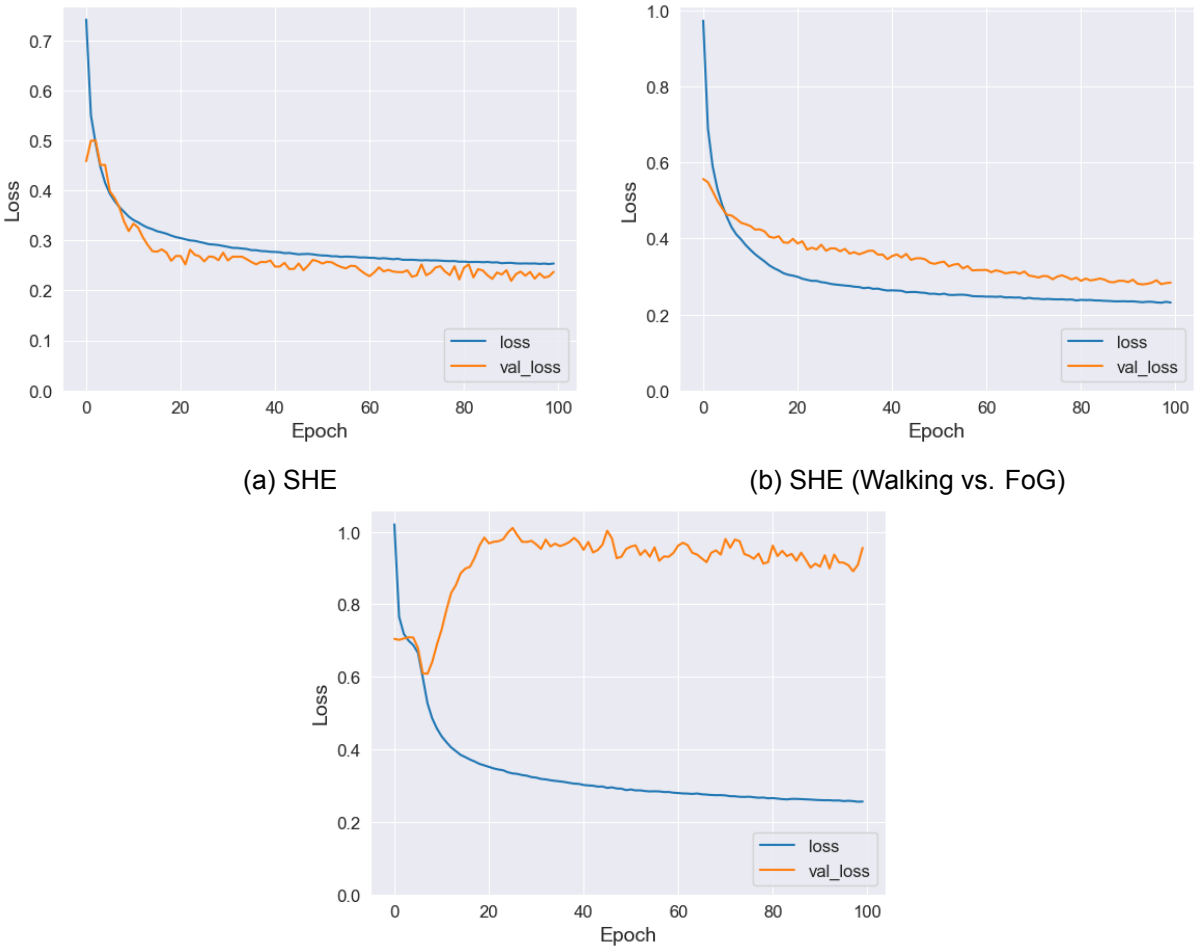


(a) SHE

(b) SHE (Walking vs. FoG)

(c) Lab

Figure 4.5: Accuracy learning curves of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset. Training and testing histories shown as 'accuracy' and 'val_accuracy' respectively.



(a) SHE

(b) SHE (Walking vs. FoG)

(c) Lab

Figure 4.6: Loss learning curves of the (a) SHE, (b) SHE (Walking vs. FoG), and (c) Lab dataset trained Multi-Headed model. Training and testing histories shown as 'loss' and 'val_loss' respectively.

5 DISCUSSION

For this research, two CNN based classification algorithms, Mono-Headed and Multi-Headed, were developed to detect FoG episodes from IMU measurements of the lower right leg in subjects with PD. Three (sub)datasets were used to train these models separately, namely the SHE, SHE (Walking vs. FoG), and Lab sets. Each set was cross validated via two schemes, 5Fold-CV and LOS-CV, to test overall performance, generalisability, and individual patient adaptability of the models. Both algorithms performed well, with average AUROCs all >91.5% for 5Fold-CV and >88.9% for LOS-CV.
These metrics are towards the higher end regarding performance, when compared with other previous studies [15, 16, 17, 25]. Furthermore, when compared directly with the recent study by Irene Heijink [21], where the same Lab dataset without data augmentation was used, an improvement of approximately >19% for the 5Fold-CV AUROC is achieved.

5.1 Interpretation of the Results

Comparing performance of both architectures on all three (sub)datasets, both architectures score similar AUROCs across both 5Fold-CV and LOS-CV. Only the Lab dataset has a noticeably improved LOS-CV AUROC score using the Multi-Headed model of 0.9331 (STD±0.0656) when compared with the Mono-Headed model of 0.9008 (STD±0.1139). From this it can be concluded that not much extra information seems to be found from extracting features from the data using multiple temporal resolutions when using the SHE dataset, while it does somewhat increase classification performance for the Lab set. An explanation for this could lie in the distribution of gait tasks being different in both datasets. The Lab dataset is made up of a multitude of gait tasks, cueing conditions, and measurement protocols, which are not homogenised across all subjects. For example, all subjects in the Hololens study [36] were asked to perform fully stationary 180° turning tasks, while all other studies contained walking tasks. This results in higher intra-class variance, for which the more simple model might not have enough trainable parameters to distinguish the complex class distribution. Yet, this can only be speculated upon with the current available data and thus should be explored further within a different study focusing on trainable parameter optimization.

Another important aspect can be gleaned from Figures 4.2 and 4.4, regarding the spread of the LOS-CV ROCs. It can be seen that, across all obtained curves, individual curves tend to be more clustered above the mean ROC and a broader spread of less curves under the mean. This shows that, while there is a minority of subjects who are hard to classify, there are comparably more subjects who the models excel at classifying. These harder subjects might have very individualised FoG presentation or their gait might be very dissimilar to the overall population. This problem of generalisation is a known problem within other studies as well regarding FoG detection [42].
Individual performance analysis of the two architectures regarding this generalisation problem shows some subtle differences when using the Lab dataset, while results obtained from both SHE datasets once again remain similar. It can be seen that there is less of a spread of subjects

under the average ROC for the Multi-Headed architecture. Figure 4.2c shows some Mono-Headed folds to be close to or on the diagonal, meaning that classifications for these subjects are mostly guesses, while Figure 4.4c shows a clear distance between the lowest performing Multi-Headed fold and the diagonal. This suggests that the Multi-Headed architecture is somewhat more able to account for subjects whose FoG and gait are more difficult to classify, thus not needing to solely guess all classifications for any subject.

While improvements of the Multi-Headed over the Mono-Headed architecture have been highlighted for the Lab dataset, it is similarly important to discuss the lack of improvement for the SHE datasets. As mentioned, all results obtained from both architectures using the SHE datasets show almost identical performance. This most likely indicates that the Mono-Headed model is already complex enough to extract features containing a high amount of information from the input data, and that no additional dominant features arise from the added temporal resolutions of the extra heads. If this is indeed the case, since all dominant features are outputs of the same head, the model will learn to focus in on this singular head and disregard the added two heads. In essence, learning the exact same way across both architectures, explaining the almost identical obtained results.

Both architectures show no observable signs of instability or erratic behaviour during training and testing, example accuracy and loss graphs of general behaviour per fold are shown in Figures 4.5 and 4.6. However, overfitting has been noticed multiple times in both architectures when using the Lab dataset. This is characterised by a significant gap between train and test curves of both accuracy and loss generally after the 7th epoch, accompanied with a drastic increase in loss for the test set. Once again referring to a previous thesis by Irene Heijink [21] where the same dataset was used, overfitting was also shown for multiple architectures. Yet, in that thesis, overfitting was more severe and learning behaviour seemed random at times. This further shows that there might be an inherent problem within the Lab dataset which may not yet be accounted for, such as differences between the individual study sets being too high.

When comparing results from multiple datasets it is important to take into their inherent differences into consideration, and what effects these differences might have on the outcomes. It has already been stated extensively that observed difficulties may have arisen from the Lab dataset's multi-study make-up. While internal protocol differences may explain at least part of these observances, per-subject FoG distribution most likely also has an effect. As stated by Pardoel et al. [15], subject-bias can easily arise when not all subjects in a set manifest FoG equally or at all during measurements. In all tested sets of this study, a small amount of subjects without FoG was present. For the SHE sets this was one subject (PD004) and for the Lab dataset three (PD105, PD107, and PD122). Full overviews of FoG distribution per patient can be found in the FoG% columns of Tables A.1-A.3 and B.1-B.3 in Appendices A and B. Thus, it can be assumed that both datasets contain a small amount of subject-based FoG bias as a consequence, yet effects of this could be enhanced due to the Lab dataset's multi-study make-up.
As stated before, the Lab dataset can be seen as a dataset consisting of four subdatasets from each study. It happens that all three subjects without manifested FoG originate from the Hololens subdataset. This dataset consists of 15 total subjects, meaning that 3 subjects without FoG is already 20% of said dataset, skewing personal-bias towards the other 80% for this set alone. Similarly, mean per-subject FoG representation is not equal over all subdatasets. While Hololens and Cinoptics are fairly balanced, with per-subject FoG means of 53% and 46% respectively, while Vibrating Socks and Pedal are much lower, 30% and 24% respectively. When combining these two factors with the observed protocol differences, it is likely that subdataset-specific confounders can be present, falsely linking protocol differences to higher or lower like-

lihood of FoG.

Lastly, comparisons should be made to results obtained from non-DL ML methods in the literature. ML, oftentimes thresholding, methods are another extensive research topic for FoG detection. These methods tend to have lower performance (results generally falling between 66.25-98.35% and 66.00-99.72% for sensitivity and specificity respectively [15]) while attaining faster processing times, thus making them potential candidates for online wearable systems. [15] Therefore, it is important to compare processing times with those of the current research. Thresholding methods require a traditional feature extraction algorithm to extract features from the data, which are then easily and quickly classified via one or more learned threshold values. Since the proposed DL method in this study only utilises a relatively simple pre-processing method, which in an online setting would consist of mainly band-pass filtration and 60 Hz resampling of a 2-second window, it is safe to assume this does not exceed times needed in threshold methods to process data and extract pre-defined features. For example, a widely used feature for current FoG detection is the freezing index [15, 43], which requires a similar pre-processing method as utilised in the current study. This means that any added processing time would solely arise from classifying the 2-second input windows with the trained model. On average these processing times are 69 ms and 79 ms for the Mono-Headed and Multi-Headed models respectively. Seeing that both classification processing times are negligible, it can be concluded that overall added computational latency of this DL technique when compared to ML thresholding techniques is minimal.

5.2 Strengths and Weaknesses

Several strengths of this research can be established, one of which being the large amount of data used to train the classification models. As mentioned, dataset size is a well-known problem in the majority of FoG research papers, which has been shown to hamper model performance. [15] This study utilised two relatively large datasets, compared to other recent studies [15], with 63.13 and 21.41 hours of usable data from 20 and 64 subjects respectively. Furthermore, the majority of these subjects experienced FoG episodes during measurements, thus producing a diverse pool of FoG data for the model to draw from.

Multiple datasets were used to show architecture performance and applicability. Data from either dataset was obtained using different IMU sensors and in different settings. Positive performance for both datasets shows that the developed architectures are versatile in use-case. Particularly, similar performance in both lab and daily life settings can be a benefit for different future works, namely further FoG research by replacing the need for video annotation or implementation in a wearable intervention or monitoring device for daily use.

Classification using only active movement data, in the SHE (Walking vs. FoG) set, was explored for both architectures as well. Here, performance was only slightly lower than when trained on the much larger complete SHE dataset. Furthermore, by downsampling nonFoG data in this manner, less synthetically generated FoG windows from data augmentation were needed to balance the set. This shows that implementing an activity detection algorithm is feasible for this dataset. By implementing such a scheme, training times could be drastically reduced without losing performance. [29]

Two cross validation schemes were applied, 5Fold-CV and LOS-CV, which give different insights into model performance and applicability. By generally scoring well on both, it has been shown that the produced models have good overall performance, generalise well, and have

acceptable to good individual patient adaptability.

The use of short windows, low latency, and a lightweight architecture paves the way for implementation in a future wearable device. Such a device is highly sought after and the main end-goal of many studies within the FoG research scope. Furthermore, usage of short windows also enables quicker reaction times for any ambulatory cueing devices, ensuring that the user receives intervention without high latency.

In addition to the strengths, there are also some weaknesses within the setup of this research. One such weakness is the aforementioned make of the Lab dataset. As mentioned, obtained results hint at the study subsets out of which the full dataset is made up to be too dissimilar. Furthermore, FoG distribution varies heavily between study subsets and per-patient. These dissimilarities could lead to the model learning confounding-factors linking FoG probability to study specific markers, such as cue-induced gait patterns. Overfitting and other learning problems have been present not only in this study, but also a previous thesis by Irene Heijink [21]. In her thesis, she came to a similar conclusion as mentioned here. This fact shows that something inherent to this dataset has, most likely, been unaccounted for and should be investigated.

While model performance was explored regarding using some sort of active movement only dataset in the SHE (Walking vs. FoG) set, this also excluded many other gait tasks as a result of how this set was produced. Preferably, only actual moments of inactivity should be excluded and thus these other gait tasks should be included. By including these gait tasks, the given data would be more well-rounded and applicable to actual FoG triggering conditions.

Another possible weakness could be the data-balancing method. Datasets were balanced according to overall dataset class distributions, rather than balancing train and test sets individually. This choice was made to not induce possible subject bias by oversampling windows of one subject more than the other. As a result, train, test, and per-subject distributions differed, thus possibly affecting classification capabilities per fold.

Data augmentation was used to upsample the FoG class. While this in itself should not be seen as a weakness, there is a choice to when and how augmentation is implemented in the pipeline. To lower computational load and training times, it was chosen to augment data during pre-processing. Since large amounts of data would be used during training, mitigating training times was of high import. Yet, by instead implementing the same data augmentation per epoch, as in the research by Camps et al. [25], augmented windows would differ each cycle and thus produced models should theoretically have better generalisability in turn.
Another potential weakness here is the choice of rotational boundaries of the data augmentation algorithm. For this research, maximum potential rotation was set to 360° for each axis, thus all possible 3D orientations of the sensor were available and randomly picked from by the algorithm. Certain of these resulting orientations would not be realistic when applied to a real world scenario. For example, the IMU's flat side (XZ-Plane) is always attached to the subject's ankle, thus an augmented 90° rotation of solely the z-axis would result in an orientation perpendicular to the skin, which would be unrealistic. Camps et al. [25] highly limited these boundaries to 30° in x- and 10° in y- and z-directions for their waist-positioned IMU, staying more realistic. Yet, by heavily limiting boundaries, generated windows will become more similar on average, lowering the amount of augmentations that can feasibly be done per window. Seeing that, theoretically, dominant positional and rotational invariant features will be extracted by the CNN [26], it is speculated that any effect of using unrealistic rotations should be minimal. Yet, utilising only realistic rotations would naturally be a safer option.

5.3 Future Recommendations

Regarding the results of this research, multiple future recommendations become apparent.
First of all, choice of model should depend highly on area of application. Both models produced comparable results, yet slight differences have shown different strengths. When seeking to apply the classifier to a possible future wearable device, the lower computational load and faster response time of the Mono-Headed model would be preferred. Conversely, when higher computational load is acceptable, the Multi-Headed model could offer slightly better personalised performance when tackling a set of highly diverse subject measurements.
Secondly, implementing a more robust activity detection algorithm such as the activity threshold method by Borzì et al. [29] could produce a more well-rounded true to life dataset for FoG classification. Another path could also be upright detection rather than activity detection, since FoG will only occur when a subject is standing or walking. Upright detection might account better for akinesia and movement initiation induced FoG, but would require multiple sensors in different areas to function.
Thirdly, applying the utilised data augmentation technique per epoch instead of during preprocessing should theoretically boost robustness of the model, thus should be explored. [25] Furthermore, exploring different kinds of data augmentation might give valuable insight into further balancing datasets without producing windows that are too similar. While the utilised rotational data augmentation technique was shown to perform as one of the best regarding FoG classification, others such as perturbation also showed promise [41]. Data augmentation techniques could also be combined to widen the range of produced windows even more.
Fourthly, exploration of different classifier modules in the architecture could heighten performance further. Both proposed architectures utilise an MLP for the binary classification part of the process, fed by features extracted by the CNN. Recent studies [15] have highlighted multiple high performance ML classifiers for FoG detection, such as Support Vector Machines, Random Forests, and AdaBoosted Decision Trees. Since MLPs are relatively simple, it is deemed likely that any of these classifiers could be adapted in place of the current MLP module to potentially further boost classification power of the architecture.

6 CONCLUSION

The aim of this research was to develop a CNN based classification model to detect FoG in PD patients using IMU data. For this, two dataset were used with 20 and 64 subjects who all experienced FoG in their daily lives. One of these datasets was further processed into a set containing only active movement data. Two architectures were developed; the lightweight Mono-Headed architecture, and the more complex Multi-Headed architecture. Both models were cross validated according to two schemes, 5Fold-CV and LOS-CV. Similar performance results in mean AUROCs were obtained, all >91.5% for 5Fold-CV and >88.9% for LOS-CV. While similar in overall performance, both models showed different strengths in deployment time and personalised performance on highly diverse subject measurement sets. Potential for deployment in a wearable setting, such as ambulatory cueing or monitoring, is deemed high. Furthermore, by utilising only one sensor, which was rated highly wearable by PD patients [16], positive adaptation of the technology is deemed likely. Future research topics and recommendations for possible improvements have been identified; choice of model should depend on area of application, improved activity or upright detection could produce more well-rounded trained models, exploration of multiple data augmentation techniques, and replacement of the MLP module by other high performance ML classifiers. Lastly, additional testing on independent datasets, introducing more subject measurements, and continued exploration of hyper parameter tuning could, as always, improve the proposed architectures further.

REFERENCES

[1] *World health organisation: Parkinson disease* [who.int/news-room/fact-sheets/detail/parkinson-disease]. (2023).

[2] Simon, D., Tanner, C., & Brundin, P. (2019). Parkinson disease epidemiology, pathology, genetics and pathophysiology. *Clinics in Geriatric Medicine*, *36*. https://doi.org/10.1016/j.cger.2019.08.002

[3] Snijders, A. H., Nijkrake, M. J., Bakker, M., Munneke, M., Wind, C., & Bloem, B. R. (2008). Clinimetrics of freezing of gait. *Movement Disorders*, *23*(S2), S468–S474. https://doi.org/https://doi-org.ezproxy2.utwente.nl/10.1002/mds.22144

[4] Zhang, W., Yang, Z., Li, H., Huang, D., Wang, L., Wei, Y., Zhang, L., Ma, L., Feng, H., Pan, J., Guo, Y., & Chan, P. (2022). Multimodal data for the detection of freezing of gait in parkinson's disease. *Scientific Data*, *9*. https://doi.org/10.1038/s41597-022-01713-8

[5] Kondo, Y., Mizuno, K., Bando, K., Suzuki, I., Nakamura, T., Hashide, S., Kadone, H., & Suzuki, K. (2022). Measurement accuracy of freezing of gait scoring based on videos. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.828355

[6] Mancini, M., Bloem, B., Horak, F., Lewis, S., Nieuwboer, A., & Nonnekes, J. (2019). Clinical and methodological challenges for assessing freezing of gait: Future perspectives. *Movement Disorders*, *34*. https://doi.org/10.1002/mds.27709

[7] Silva de Lima, A. L., Evers, L. J. W., Hahn, T., Bataille, L., Hamilton, J. L., Little, M. A., Okuma, Y., Bloem, B. R., & Faber, M. J. (2017). Freezing of gait and fall detection in parkinson's disease using wearable sensors: A systematic review. *Journal of Neurology*, *264*, 1642–1654. https://api.semanticscholar.org/CorpusID:3626185

[8] Nutt, J., Bloem, B., Giladi, N., Hallett, M., Horak, F., & Nieuwboer, A. (2011). Freezing of gait: Moving forward on a mysterious clinical phenomenon. *Lancet neurology*, *10*, 734–44. https://doi.org/10.1016/S1474-4422(11)70143-0

[9] Heremans, E., Nieuwboer, A., & Vercruysse, S. (2013). Freezing of gait in parkinson's disease: Where are we now? *Current neurology and neuroscience reports*, *13*, 350. https://doi.org/10.1007/s11910-013-0350-7

[10] Pelicioni, P. H. S., Menant, J. C., Latt, M. D., & Lord, S. R. (2019). Falls in parkinson's disease subtypes: Risk factors, locations and circumstances. *International journal of environmental research and public health, 16(12), 2216.* https://doi.org/10.3390/ijerph16122216

[11] Ghielen, I., Koene, P., Twisk, J. W., Kwakkel, G., van den Heuvel, O. A., & van Wegen, E. E. (2020). The association between freezing of gait, fear of falling and anxiety in parkinson's disease: A longitudinal analysis [PMID: 32552383]. *Neurodegenerative Disease Management*, *10*(3), 159–168. https://doi.org/10.2217/nmt-2019-0028

[12] Gao, C., Liu, J., Tan, Y., & Chen, S. (2020). Freezing of gait in parkinson's disease: Pathophysiology, risk factors and treatments. *Translational Neurodegeneration*, *9*. https://doi.org/10.1186/s40035-020-00191-5

[13] Das, J., Vitório, R., Butterfield, A., Morris, R., Graham, L., Barry, G., Mcdonald, C., Walker, R., Mancini, M., & Stuart, S. (2022). Visual cues for turning in parkinson's disease. *Sensors*, *22*. https://doi.org/10.3390/s22186746

[14] Ginis, P., Heremans, E., Ferrari, A., Bekkers, E., Canning, C., & Nieuwboer, A. (2017). External input for gait in people with parkinson's disease with and without freezing of gait:

One size does not fit all. *Journal of Neurology*, *264*. https://doi.org/10.1007/s00415-017-8552-6

[15] Pardoel, S., Kofman, Nantel, J., & Lemaire, E. (2019). Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: A review. *Sensors*, *19*, 5141. https://doi.org/10.3390/s19235141

[16] O'Day, J., Lee, M., Seagers, K., Hoffman, S., Jih-Schiff, A., Kidziński, Ł., Delp, S., & Brontë-Stewart, H. (2022). Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of NeuroEngineering and Rehabilitation*, *19*. https://doi.org/10.1186/s12984-022-00992-x

[17] Sigcha, L. F., Costa, N., Pavón, I., Costa, S., Arezes, P., Lopez Navarro, J. M., & Arcas, G. (2020). Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors. *Sensors*, *20*, 1895. https://doi.org/10.3390/s20071895

[18] Mazilu, S., Calatroni, A., Gazit, E., Mirelman, A., Hausdorff, J. M., & Tröster, G. (2015). Prediction of freezing of gait in parkinson's from physiological wearables: An exploratory study. *IEEE Journal of Biomedical and Health Informatics*, *19*(6), 1843–1854. https://doi.org/10.1109/JBHI.2015.2465134

[19] Li, B., Li, Y., Sun, Y., Yang, X., Zhou, X., & Yao, Z. (2023). A monitoring method of freezing of gait based on multimodal fusion. *Biomedical Signal Processing and Control*, *82*, 104589. https://doi.org/https://doi.org/10.1016/j.bspc.2023.104589

[20] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J. I., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, *8*. https://api.semanticscholar.org/CorpusID:232434552

[21] Heijink, I. B. (2022). *Detection of freezing of gait in patients with parkinson's disease using a deep learning approach*.

[22] Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). Minirocket: A very fast (almost) deterministic transform for time series classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 248–257. https://doi.org/10.1145/3447548.3467231

[23] Ismail Fawaz, H., Lucas, B., & Forestier, G. (2020). Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*. https://doi.org/https://doi.org/10.1007/s10618-020-00710-y

[24] Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.*, *35*(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

[25] Camps, J., Samà, A., Martín, M., Rodríguez-Martín, D., Pérez-López, C., Moreno Arostegui, J. M., Cabestany, J., Català, A., Alcaine, S., Mestre, B., Prats, A., Crespo-Maraver, M. C., Counihan, T. J., Browne, P., Quinlan, L. R., Laighin, G. Ó., Sweeney, D., Lewy, H., Vainstein, G., … Rodríguez-Molinero, A. (2018). Deep learning for freezing of gait detection in parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge-Based Systems*, *139*, 119–131. https://doi.org/https://doi.org/10.1016/j.knosys.2017.10.017

[26] Saha, S. (2018). *A comprehensive guide to convolutional neural networks — the eli5 way*. Towards Data Science. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[27] Browniee, J. (2020). *A gentle introduction to the rectified linear unit (relu)*. Machine Learning Mastery. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

[28] Huber, J. (2020). *Batch normalization in 3 levels of understanding: What do we know about it so far: From a 30 seconds digest to a comprehensive guide*. Towards Data Science. https://towardsdatascience.com/batch-normalization-in-3-levels-of-understanding-14c2da90a338#b93c

[29] Borzí, L., Sigcha, L. F., Rodríguez-Martín, D., & Olmo, G. (2022). Real-time detection of freezing of gait in parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. *Artificial Intelligence in Medicine*, *135*, 102459. https://doi.org/10.1016/j.artmed.2022.102459

[30] Furnieles, G. (2022). *Sigmoid and softmax functions in 5 minutes: The math behind two of the most used activation functions in machine learning*. Towards Data Science. https://towardsdatascience.com/sigmoid-and-softmax-functions-in-5-minutes-f516c80ea1f9

[31] Ajagekar, A. (2021). *Adam*. Cornell University. https://optimization.cbe.cornell.edu/index.php?title=Adam

[32] Leikin, I. (n.d.). *Understanding binary cross-entropy and log loss for effective model monitoring*. Aporia. https://www.aporia.com/learn/understanding-binary-cross-entropy-and-log-loss-for-effective-model-monitoring/

[33] Gupta, P. (2017). *Cross-validation in machine learning*. Towards Data Science. https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

[34] Lyashenko, V., & Jha, A. (2023). *Cross-validation in machine learning: How to do it right*. MLOps. https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

[35] Janssen, S., Bolte, B., Nonnekes, J., Bittner, M., Bloem, B., Heida, T., Zhao, Y., & Wezel, R. (2017). Usability of three-dimensional augmented visual cues delivered by smart glasses on (freezing of) gait in parkinson's disease. *Frontiers in Neurology*, *8*. https://doi.org/10.3389/fneur.2017.00279

[36] Janssen, S., de Ruyter van Steveninck, J., Salim, H., Cockx, H., Bloem, B., Heida, T., & Wezel, R. (2020). Citation: 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. https://doi.org/10.3389/fneur.2020.00185

[37] Janssen, S., Heijs, J., Bittner, M., Droog, E., Bloem, B., Wezel, R., & Heida, C. (2021). Visual cues added to a virtual environment paradigm do not improve motor arrests in parkinson's disease. *Journal of neural engineering*, *18*. https://doi.org/10.1088/1741-2552/abe356

[38] Klaver, E., Vugt, J., Bloem, B., Wezel, R., Nonnekes, J., & Tjepkema-Cloostermans, M. (2023). Good vibrations: Tactile cueing for freezing of gait in parkinson's disease. *Journal of Neurology*, *270*, 1–9. https://doi.org/10.1007/s00415-023-11663-9

[39] Moore, S. T., MacDougall, H. G., & Ondo, W. G. (2008). Ambulatory monitoring of freezing of gait in parkinson's disease. *Journal of Neuroscience Methods*, *167*(2), 340–348. https://doi.org/https://doi.org/10.1016/j.jneumeth.2007.08.023

[40] Mazilu, S., Hardegger, M., Zhu, Z., Roggen, D., Troester, G., Plotnik, M., & Hausdorff, J. (2012). Online detection of freezing of gait with smartphones and machine learning techniques. https://doi.org/10.4108/icst.pervasivehealth.2012.248680

[41] Um, T. T., Pfister, F. M. J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., & Kulić, D. (2017). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 216–220. https://doi.org/10.1145/3136755.3136817

[42] Mikos, V., Heng, C.-H., Tay, A., Chia, N. S. Y., Koh, K. M. L., Tan, D. M. L., & Au, W. L. (2017). Real-time patient adaptivity for freezing of gait classification through semi-supervised neural networks. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 871–876. https://doi.org/10.1109/ICMLA.2017.00-46

[43] Cockx, H., Nonnekes, J., Bloem, B., van Wezel, R., Cameron, I., & Wang, Y. (2023). Dealing with the heterogeneous presentations of freezing of gait: How reliable are the

freezing index and heart rate for freezing detection? *Journal of neuroengineering and rehabilitation*, *20*. https://doi.org/10.1186/s12984-023-01175-y

A MONO-HEADED ARCHITECTURE: LOS-CV METRICS

Table A.1: Per subject results of each LOS-CV fold for the SHE dataset, using the Mono-Headed Architecture. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
PD004	0.9794	nan	0.9794	0.0	nan	0.0
PD005	0.9323	0.8632	0.9442	0.7887	0.9496	14.63
PD008	0.955	0.3269	0.9577	0.0589	0.6235	0.43
PD015	0.9415	0.2841	0.9447	0.0444	0.9088	0.48
PD016	0.9395	0.4261	0.9539	0.2775	0.7431	2.73
PD019	0.9174	0.9042	0.9202	0.7915	0.9636	17.34
PD021	0.8414	0.8052	0.8898	0.8531	0.9293	57.19
PD028	0.9475	0.7662	0.9625	0.6901	0.9712	7.63
PD032	0.9501	0.5091	0.959	0.2882	0.9236	1.99
PD034	0.9728	0.5226	0.9827	0.4521	0.9275	2.14
PD038	0.9136	0.8762	0.9309	0.865	0.9648	31.59
PD039	0.949	0.7134	0.9591	0.5345	0.9468	4.1
PD040	0.9031	0.9075	0.8992	0.8971	0.968	46.56
PD043	0.7727	0.5809	0.956	0.7141	0.9024	48.88
PD044	0.8744	0.7616	0.9525	0.8323	0.9454	40.93
PD045	0.7754	0.7779	0.7725	0.7878	0.8541	53.6
PD046	0.9183	0.7825	0.9445	0.7561	0.9556	16.18
PD047	0.9364	0.9232	0.9398	0.8542	0.9784	20.17
PD048	0.9486	0.7921	0.9656	0.7508	0.9729	9.77
PD049	0.8502	0.6899	0.9605	0.789	0.9478	40.74
Average	0.9109	0.6954	0.9387	0.6013	0.9145	-
STD	0.0577	0.1922	0.0444	0.2951	0.0867	-

Table A.2: Per subject results of each LOS-CV fold for the SHE (Walking vs. F0G) dataset, using the Mono-Headed Architecture. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
PD004	0.9317	nan	0.9317	0.0	nan	0.0
PD005	0.8433	0.8612	0.8382	0.7112	0.9317	22.4
PD008	0.2741	0.6731	0.2706	0.0159	0.5139	0.87
PD015	0.8831	1.0	0.8811	0.2176	0.9841	1.63
PD016	0.7348	0.8908	0.7242	0.298	0.8274	6.32
PD019	0.9012	0.9449	0.8816	0.855	0.9749	30.85
PD021	0.9547	0.9764	0.8829	0.9706	0.9826	76.78
PD028	0.9264	0.9143	0.9279	0.7356	0.9806	11.19

PD032	0.8362	0.9268	0.8308	0.3868	0.962	5.57
PD034	0.8503	0.9729	0.8427	0.4324	0.9678	5.86
PD038	0.8405	0.9855	0.6491	0.8755	0.867	56.9
PD039	0.8652	0.9528	0.851	0.6628	0.9695	13.91
PD040	0.9505	0.9914	0.7791	0.97	0.9658	80.75
PD043	0.9141	0.9958	0.6667	0.9457	0.9264	75.17
PD044	0.9437	0.9833	0.8603	0.9595	0.9797	67.8
PD045	0.9361	0.9995	0.1384	0.9666	0.5931	92.64
PD046	0.9562	0.9876	0.3069	0.9773	0.5901	95.39
PD047	0.8718	0.9376	0.8456	0.8063	0.9558	28.46
PD048	0.852	0.9741	0.8024	0.7916	0.9618	28.86
PD049	0.9301	0.989	0.8332	0.9462	0.9688	62.22
Average	0.8598	0.9451	0.7372	0.6762	0.8896	-
STD	0.1448	0.0745	0.2232	0.3217	0.1463	-

Table A.3: Per subject results of each LOS-CV fold for the Lab dataset, using the Mono-Headed Architecture. The first two letters of each subject tag indicate which study they originate from; ID are Cinoptics and Pedal, PD is Hololens, and VS is Vibrating Socks. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
ID01	0.9028	0.9375	0.9002	0.5714	0.955	6.91
ID02	0.887	0.95	0.8857	0.2585	0.9776	2.07
ID03	0.9083	0.9891	0.0879	0.9515	0.744	91.03
ID04	0.5652	0.9544	0.4886	0.4194	0.7418	16.45
ID05	0.9291	0.9583	0.9274	0.6079	0.9926	5.73
ID05	0.9461	0.9766	0.8202	0.9669	0.964	80.5
ID06	0.9503	1.0	0.9478	0.6564	0.9952	4.75
ID07	0.866	0.9836	0.7062	0.8943	0.951	57.63
ID07	0.7345	0.95	0.7221	0.2804	0.9317	5.45
ID09	0.8472	0.9562	0.7063	0.8758	0.9186	56.37
ID10	0.8205	0.7656	0.8229	0.27	0.862	4.34
ID12	0.9355	0.9444	0.935	0.6071	0.9861	5.27
ID13	0.8665	0.8284	0.9133	0.8725	0.9396	55.17
ID14	0.9257	0.9014	0.9346	0.8671	0.9771	26.9
ID15	0.8067	0.9392	0.6929	0.8179	0.9056	46.21
ID16	0.813	0.9122	0.7035	0.8366	0.9081	52.47
ID18	0.8414	0.945	0.7154	0.8673	0.9291	54.86
ID19	0.9749	0.9887	0.9026	0.9851	0.9798	84.03
ID19	0.9366	0.9154	0.9547	0.9299	0.9761	45.92
ID20	0.9201	0.872	0.9258	0.6976	0.9702	10.57
ID23	0.8354	0.872	0.7845	0.8602	0.9268	58.1
ID26	0.831	0.8854	0.8257	0.4809	0.9458	8.84
ID27	0.9362	0.9561	0.8513	0.9604	0.9717	80.98
ID28	0.9205	0.8906	0.9689	0.9326	0.9824	61.78
ID29	0.9193	0.9252	0.9001	0.9463	0.9716	76.76
PD02	0.9565	0.9989	0.0	0.9778	0.6061	95.75
PD05	0.7994	0.8167	0.7782	0.8178	0.8868	55.13
PD08	0.9585	0.9652	0.9423	0.9706	0.9884	70.85

PD105	0.3678	nan	0.3678	0.0	nan	0.0
PD107	0.6904	nan	0.6904	0.0	nan	0.0
PD122	0.9759	nan	0.9759	0.0	nan	0.0
PD18	0.7115	0.7656	0.6931	0.5731	0.8561	25.3
PD19	0.9896	0.995	0.6311	0.9947	0.9541	98.51
PD27	0.9265	0.9676	0.0206	0.9618	0.4913	95.67
PD76	0.5663	0.815	0.4274	0.5739	0.5879	35.84
PD84	0.8697	0.875	0.8689	0.6364	0.9444	13.03
PD87	0.8079	0.7547	0.8631	0.8	0.8747	50.9
PD92	0.9853	0.9878	0.0	0.9926	0.794	99.75
PD95	0.7923	0.8401	0.6616	0.8556	0.8683	73.23
PD99	0.9274	0.994	0.387	0.9606	0.9026	89.04
VS01	0.9577	0.7337	0.9905	0.8157	0.9818	12.77
VS03	0.4575	0.7125	0.4442	0.1147	0.7447	4.93
VS05	0.4545	0.9507	0.2702	0.4857	0.7355	27.09
VS06	0.7309	0.997	0.118	0.8378	0.6883	69.73
VS09	0.9961	0.9632	0.9981	0.9668	0.9995	5.92
VS10	0.9181	0.934	0.9152	0.7818	0.9825	15.71
VS14	0.9426	0.9751	0.8586	0.9607	0.9834	72.07
VS15	0.8189	0.9334	0.8021	0.5683	0.9667	12.77
VS17	0.475	0.95	0.41	0.3035	0.8542	12.04
VS18	0.6833	0.9446	0.5104	0.7038	0.9049	39.83
VS19	0.2676	0.8173	0.236	0.1082	0.6174	5.44
VS23	0.8201	0.9516	0.7999	0.5838	0.9609	13.26
VS24	0.9226	0.8665	0.934	0.7901	0.961	16.8
VS26	0.8885	0.9245	0.7773	0.9261	0.9299	75.56
VS29	0.8716	0.6655	0.9925	0.7931	0.9515	36.97
VS30	0.8818	0.9256	0.8743	0.6956	0.9713	14.6
VS31	0.9285	0.9597	0.8835	0.9408	0.9834	59.14
VS32	0.956	0.8914	0.9879	0.9305	0.9817	33.07
VS33	0.9534	0.9062	0.9543	0.4099	0.9726	1.78
VS34	0.7807	0.8976	0.6703	0.7991	0.9119	48.59
VS37	0.8772	0.9117	0.7493	0.9212	0.9414	78.76
VS38	0.6391	0.861	0.5558	0.5656	0.7965	27.3
VS39	0.9739	0.9352	0.9783	0.8783	0.9934	10.06
VS40	0.9449	0.8529	0.9752	0.8846	0.9765	24.75
Average	0.8326	0.9088	0.719	0.7425	0.9008	-
STD	0.159	0.0761	0.272	0.2335	0.1139	-

B MULTI-HEADED ARCHITECTURE: LOS-CV METRICS

Table B.1: Per subject results of each LOS-CV fold for the SHE dataset, using the Multi-Headed Architecture. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
PD004	0.9607	nan	0.9607	0.0	nan	0.0
PD005	0.9382	0.8841	0.9474	0.8071	0.9546	14.63
PD008	0.9767	0.0	0.9809	0.0	0.5492	0.43
PD015	0.8389	0.983	0.8382	0.0552	0.9683	0.48
PD016	0.9267	0.4507	0.9401	0.2512	0.7628	2.73
PD019	0.8896	0.9401	0.879	0.7471	0.9669	17.34
PD021	0.8529	0.8433	0.8657	0.8677	0.9295	57.19
PD028	0.9244	0.9228	0.9245	0.6507	0.9755	7.63
PD032	0.9383	0.9177	0.9387	0.3711	0.9796	1.99
PD034	0.9617	0.634	0.9689	0.4152	0.932	2.14
PD038	0.8738	0.8793	0.8713	0.8149	0.9299	31.59
PD039	0.9548	0.6934	0.966	0.5571	0.9297	4.1
PD040	0.8945	0.9406	0.8542	0.8925	0.9482	46.56
PD043	0.7882	0.6483	0.9221	0.7495	0.9007	48.88
PD044	0.8649	0.7641	0.9348	0.8224	0.9253	40.93
PD045	0.7607	0.801	0.7141	0.782	0.8164	53.6
PD046	0.9092	0.8458	0.9214	0.7509	0.9474	16.18
PD047	0.9358	0.8866	0.9482	0.8478	0.9736	20.17
PD048	0.9404	0.8649	0.9485	0.7392	0.9706	9.77
PD049	0.8773	0.7931	0.9353	0.8405	0.949	40.74
Average	0.9004	0.7733	0.913	0.5981	0.911	-
STD	0.0566	0.2227	0.0604	0.297	0.1006	-

Table B.2: Per subject results of each LOS-CV fold for the SHE (Walking vs. FoG) dataset, using the Multi-Headed Architecture. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
PD004	0.9478	nan	0.9478	0.0	nan	0.0
PD005	0.8648	0.8712	0.863	0.7427	0.9386	22.4
PD008	0.2854	0.6346	0.2823	0.0152	0.5374	0.87
PD015	0.8921	1.0	0.8903	0.2316	0.9782	1.63
PD016	0.7492	0.8627	0.7416	0.303	0.8626	6.32
PD019	0.9044	0.9435	0.887	0.8589	0.9685	30.85
PD021	0.9561	0.9725	0.9018	0.9715	0.983	76.78
PD028	0.9259	0.9382	0.9244	0.7393	0.9766	11.19

PD032	0.8725	0.9085	0.8704	0.4428	0.9549	5.57
PD034	0.8687	0.9789	0.8618	0.4663	0.9728	5.86
PD038	0.853	0.9832	0.681	0.8838	0.8601	56.9
PD039	0.8822	0.9434	0.8723	0.6903	0.9699	13.91
PD040	0.9489	0.9922	0.7676	0.9691	0.9613	80.75
PD043	0.9104	0.9958	0.6518	0.9435	0.9504	75.17
PD044	0.9436	0.9825	0.8617	0.9594	0.9744	67.8
PD045	0.9331	0.9997	0.0956	0.9651	0.5463	92.64
PD046	0.9526	0.9837	0.3069	0.9754	0.5757	95.39
PD047	0.8803	0.9343	0.8588	0.8163	0.9563	28.46
PD048	0.855	0.9717	0.8077	0.7946	0.9604	28.86
PD049	0.9301	0.9898	0.8318	0.9463	0.9755	62.22
Average	0.8678	0.9414	0.7453	0.6858	0.8896	-
STD	0.142	0.0826	0.2321	0.3176	0.1496	-

Table B.3: Per subject results of each LOS-CV fold for the Lab dataset, using the Multi-Headed Architecture. The first two letters of each subject tag indicate which study they originate from; ID are Cinoptics and Pedal, PD is Hololens, and VS is Vibrating Socks. FoG% indicates individual FoG distribution for that subject's data.

Subject	Accuracy	Sensitivity	Specificity	F1-Score	AUROC	FoG%
ID01	0.915	0.9688	0.9111	0.6118	0.9674	6.91
ID02	0.9243	0.975	0.9232	0.3482	0.9876	2.07
ID03	0.8569	0.8544	0.8821	0.9157	0.9501	91.03
ID04	0.6025	0.9235	0.5392	0.4333	0.8553	16.45
ID05	0.9586	0.9444	0.9595	0.7234	0.9883	5.73
ID05	0.9478	0.9817	0.808	0.968	0.9588	80.5
ID06	0.954	0.9844	0.9525	0.6702	0.9908	4.75
ID07	0.8853	0.9661	0.7754	0.9066	0.9498	57.63
ID07	0.757	0.9375	0.7466	0.2959	0.9318	5.45
ID09	0.8571	0.9119	0.7862	0.8779	0.9071	56.37
ID10	0.8604	0.8125	0.8626	0.3355	0.8982	4.34
ID12	0.9641	0.9583	0.9644	0.738	0.9894	5.27
ID13	0.8763	0.8205	0.9449	0.8798	0.9446	55.17
ID14	0.9305	0.8389	0.9642	0.8666	0.9791	26.9
ID15	0.8288	0.9142	0.7554	0.8315	0.9307	46.21
ID16	0.833	0.8493	0.8151	0.8422	0.9152	52.47
ID18	0.868	0.9328	0.7891	0.8857	0.9536	54.86
ID19	0.9525	0.9546	0.9415	0.9713	0.9799	84.03
ID19	0.9363	0.9314	0.9405	0.9307	0.9692	45.92
ID20	0.9394	0.8567	0.9492	0.7493	0.9742	10.57
ID23	0.8484	0.8488	0.8478	0.8668	0.9309	58.1
ID26	0.877	0.9062	0.8742	0.5659	0.9664	8.84
ID27	0.9261	0.9416	0.8601	0.9538	0.9655	80.98
ID28	0.9202	0.8932	0.9638	0.9326	0.9803	61.78
ID29	0.9287	0.9329	0.915	0.9526	0.9769	76.76
PD02	0.954	0.9751	0.4793	0.976	0.9333	95.75
PD05	0.8545	0.825	0.8908	0.8621	0.9294	55.13
PD08	0.9529	0.9541	0.95	0.9663	0.9866	70.85

PD105	0.3808	nan	0.3808	0.0	nan	0.0
PD107	0.6995	nan	0.6995	0.0	nan	0.0
PD122	0.9845	nan	0.9845	0.0	nan	0.0
PD18	0.8617	0.7031	0.9153	0.72	0.8733	25.3
PD19	0.9886	0.9948	0.582	0.9942	0.95	98.51
PD27	0.8997	0.9334	0.1551	0.9468	0.8165	95.67
PD76	0.7706	0.95	0.6704	0.748	0.9178	35.84
PD84	0.9511	0.9219	0.9555	0.831	0.9672	13.03
PD87	0.7779	0.5991	0.9633	0.733	0.8328	50.9
PD92	0.9809	0.9821	0.4706	0.9903	0.9473	99.75
PD95	0.81	0.8164	0.7925	0.8629	0.8763	73.23
PD99	0.924	0.9853	0.4263	0.9585	0.8867	89.04
VS01	0.9611	0.7663	0.9897	0.8343	0.978	12.77
VS03	0.5752	0.7375	0.5668	0.1462	0.823	4.93
VS05	0.4854	0.9449	0.3146	0.4987	0.8638	27.09
VS06	0.742	0.9915	0.1672	0.8427	0.6995	69.73
VS09	0.9996	0.9926	1.0	0.9963	0.9996	5.92
VS10	0.957	0.9245	0.9631	0.8711	0.9897	15.71
VS14	0.9541	0.9841	0.8766	0.9686	0.9837	72.07
VS15	0.9119	0.8596	0.9196	0.7137	0.9589	12.77
VS17	0.6162	0.8196	0.5884	0.3396	0.8519	12.04
VS18	0.7762	0.8491	0.7279	0.7513	0.9017	39.83
VS19	0.2561	0.9968	0.2136	0.1272	0.6845	5.44
VS23	0.8973	0.9093	0.8955	0.7014	0.9702	13.26
VS24	0.9298	0.8184	0.9523	0.7967	0.9626	16.8
VS26	0.881	0.8583	0.951	0.9159	0.9748	75.56
VS29	0.874	0.6598	0.9996	0.7947	0.9587	36.97
VS30	0.9383	0.8877	0.947	0.8078	0.9776	14.6
VS31	0.927	0.9722	0.8617	0.9403	0.9803	59.14
VS32	0.955	0.8891	0.9875	0.9289	0.983	33.07
VS33	0.9875	0.8906	0.9892	0.717	0.9835	1.78
VS34	0.8001	0.9172	0.6894	0.8168	0.9326	48.59
VS37	0.878	0.9247	0.7046	0.9227	0.9177	78.76
VS38	0.6559	0.8174	0.5953	0.5646	0.8136	27.3
VS39	0.9739	0.8519	0.9876	0.8679	0.9956	10.06
VS40	0.9413	0.8242	0.9798	0.8742	0.9764	24.75
Average	0.8565	0.8945	0.7946	0.7403	0.9331	-
STD	0.1446	0.0835	0.2173	0.2632	0.0656	-