Detecting Freezing of Gait Using Wearable Sensors and Machine Learning: Exploring Ternary Freezing of Gait Classification

by

Andrew Hart

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical & Mechatronics Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This work focuses on Parkinson's disease (PD), a neurodegenerative disease characterized by the production of Lewy bodies in the brain, resulting in the degeneration of dopaminergic nigrostriatal neurons. A common and debilitating symptom of PD is Freezing of Gait (FoG), which is described as a sudden, episodic inability to make forward progress while walking despite the intention to do so. FoG can lead to falls and difficulties in everyday tasks, especially mobility. Conventional PD treatments have a variable impact on mitigating FoG due to large heterogeneity within the freezing population, necessitating active monitoring of an individual's FoG severity. This study aims to aid the development of active FoG severity monitoring using wearable sensors and machine learning. Specifically, it explores the ternary (3-class) domain of FOG classification (akinetic, kinetic, and no FoG), which has not been extensively studied before. Specific objectives of this thesis comprises of: identifying suitable datasets, selecting and optimizing machine learning models, evaluating model performance on participants, and identifying potential applications based on observed results.

Two datasets were considered for this study, including the Sydney dataset collected by Goh et al. at the University of Sydney in Australia, and the publicly-available MJFF dataset comprising multiple collections of data from various groups. The Sydney dataset consists of 10 participants completing the Ziegler protocol in their "ON" and "OFF" medication states while equipped with a tri-axial inertial measurement unit (IMU) on their sternum, lumbar, and bilateral feet. Throughout this dataset, there was a total of 24.9% of the time spent in an akinetic freeze and 8.87% of the time spent in a kinetic freeze. As for the MJFF dataset, it was comprised of 100 participants completing a similar Ziegler protocol and an alternative DeFOG protocol in the two medication states with a lumbar tri-axial accelerometer. In total, there were 833 trials for the Zeigler protocol in this dataset, and 91 trials for the DeFOG protocol, combining to produce a total of 1.47% of the time spent in an akinetic freeze and 12.39% of the time spent in a kinetic freeze states.

For classification models, a total of seven architectures were considered, including six classical models and one deep network model. The classical models received input in the form of feature vectors, whereas the deep model utilized frequency domain signals along with a convolutional network backbone to extract information. The features included in this study were selected from establishing an initial pool, then trimming the included features down using common feature engineering techniques such as Kendalls correlation, and Minimum Redundancy - Maximum Relevance (mRMR). Additionally, all models went through a randomized grid search for the optimal hyperparameters and architecture parameters to optimize performance on the utilized datasets.

Testing the models with the participant data in the Sydney dataset revealed that all classical models and the deep network model encountered challenges in ternary FoG classification compared to results in the current literature. While some models performed well for a subset of participants, mainly severe freezers, the majority of the classifiers struggled to accurately label ternary FoG bouts with many F1-scores falling below 40%. The top-performing classical model, logistic regression (LR), faced difficulties in classifying kinetic freezing and temporal accuracy. It was theorized these difficulties arose due to limited frequency domain features in the final feature set, and limited information about neighbouring windows when making inferences. While the deep model also struggled with correctly classifying the timing of the bout, to a larger extent, it had trouble differentiating between akinetic and kinetic freezing. This drop in performance is likely attributable to freeze states not achieving steady state, and/or the large heterogeneity within the population producing in manifesting akinetic and kinetic freezing (e.g., some akinetic freezes might have movement, while others are purely akinetic with no movement at all).

When FoG onsets and offsets were not considered, both models demonstrated better performance in classifying severity, with the LR model predicting correct severity for seven out of ten individuals and achieving an F1-score of 76% in akinetic freezing and correctly predicting six out of ten individuals and achieving an F1-score of 60% in kinetic freezing. The deep model correctly classified the combined total severity (akinetic and kinetic percentages combined) for seven out of ten individuals and achieved an F1-score of 58%.

The findings of this thesis indicate that existing models face challenges in automatically detecting ternary FoG labels. Further exploration of feature pools and architectures is warranted to enhance performance in free-living applications. Post-calibration techniques on model outputs or combining models in a majority voting system are recommended. Ultimately, this study suggests that the current use of ternary FoG classification may be better suited for severity estimates or as an annotation tool for clinicians, rather than a gold standard for free-living labels. More specifically, the models could be used to provide severity estimates in free-living conditions. These estimates could be later combined with in-clinic visits to gain a deeper understanding of an individual's disease progression. Alternatively, actual FoG bout classification can serve as a tool to expedite annotators by flagging areas of interest prior to a manual confirmation process.

Acknowledgements

I would like to thank my supervisor, Professor James Tung, for his extensive support and guidance through my MASc journey.

I would also like to thank my thesis committee, the NRE lab, the NiMBaL lab, and ONDRI group members.

Lastly, I would like to express my gratitude to all individuals who participated in the included studies.

Dedication

This is dedicated to my dearest Stephanie and Bailey.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADT AdaBoosted decision tree 13, 14, 29, 33, 54, 58, 82

ANOVA analysis of variance 11

CNN convolutional neural network 15

DBS deep brain stimulation 2

ECG electrocardiogram 8, 10

FFT fast Fourier transform 25, 26

FoG freezing of gait 1–19, 21, 22, 24, 27, 29, 38, 41, 42, 50–52, 58–63, 65, 66, 69, 70, 72–74, 78

IMU inertial measurement unit 8–10, 19, 25, 27, 28, 41, 44

KNN k-nearest neighbours 13, 29, 32, 54, 58, 68, 82

LOSO leave one subject out 14, 40, 41, 51

LR logistic regression 13, 29, 30, 34, 54, 58, 60, 61, 63, 64, 66–69, 71, 85

MDS-Unified Movement Disorder Society-Unified 17

mRMR minimum Redundancy - Maximum Relevance 11, 28

NN neural network 13, 14, 29, 34, 68, 82

OVR one versus rest 30, 31

PD Parkinson's disease 1–4, 6–8, 10, 17

RF random forest 13, 14, 29, 30, 33, 68, 82

RNN recurrent neural network 14

SVM support vector machine 13, 14, 29, 31, 54, 58, 59, 68, 82

TUG timed up and go 16

Chapter 1

Introduction

1.1 Parkinson's Disease

Parkinson's disease (PD) is a neurodegenerative disease that can be characterized by the production of Lewy bodies in the brain that cause degeneration of dopaminergic nigrostriatal neurons [11]. Considering PD mainly affects older adults and is the second most common neurodegenerative disease worldwide, PD will inevitably become more prevalent as current larger generations grow older [12].

1.1.1 Freezing of Gait

One of the common and debilitating symptoms of PD is freezing of gait (FoG), defined as a "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk" [13, 14]. Often people with FoG describe it as the feeling of your feet being glued to the floor despite the intention to make your next step. This makes FoG a concerning symptom for the PD population since this lack of forward progression (despite intention) can lead to falls and be debilitating to everyday mobility tasks [15]. In recent studies, FoG events have been viewed to belong to two main subtypes: 1) akinetic, where the individual exhibits a complete lack of movement during the freeze; 2) kinetic, where the individual exhibits tremor in the lower limbs or very short shuffling steps during the freeze [13].

At the time of writing, the severity of FoG is measured in-clinic dominantly by gait analysis and/or questionnaires [16, 10, 17, 18, 19, 20]. Gait analysis can be completed in real-time, or with the assistance of video annotations. Video annotations provide the benefit of a closer look at an individual's gait patterns, but are often time-consuming for experts and can introduce inter-rater reliability challenges. Regardless of the method for analysis, the gait features are often used in a rating system such as the Ziegler severity score that helps to fit a certain trial onto an established scale [16, 10]. A large proportion of these rating systems utilize the subtype of freeze as a dominating metric (e.g., kinetic, or akinetic) when binning the severity [17, 18].

As for questionnaires, this form of assessment can help reduce the time commitment but is vulnerable to subjectivity among individuals. The primary outcome of questionnaires can also vary heavily, with some such as the NFOG primarily focused on the onset of FoG which in turn struggles with severity [21]. While other questionnaires, such as the C-FOG, aim to highlight the heterogeneity of FoG triggers within the target population [22].

1.2 Motivation

Individuals that experience FoG can seek aid through conventional PD treatments including deep brain stimulation (DBS), medication (e.g., Levodopa), or gait therapy [23, 24]. However, due to the large amount of heterogeneity within the FoG population, understanding the benefits of specific treatment plans throughout an individual's disease progression requires active monitoring of their symptom severity. Monitoring of an individual's FoG severity will help characterize the response to specific treatments and will aid in making adjustments to improve quality of life [23, 24, 25].

A crucial step to enable active monitoring is the development of automation tools that can streamline the current assessment techniques that are done manually. This development aims to include both in-clinic and remote FoG assessments. Automatic modelling, with current machine learning classification and wearable sensor technologies, can be utilized for feats such as FoG classification. In current literature, this topic has been explored extensively in the binary domain, investigating no FoG vs FoG labels only [26]. At the time of writing, little to no work has been published extrapolating this solution space to the ternary domain to break FoG labels into separate akinetic and kinetic freezing types, along with no FoG.

As mentioned, manifestation of FoG is highly heterogeneous across the population. This work proposes a ternary FoG classification approach towards a more granular individual severity assessment. In clinical assessments, severity estimates often include FoG subtype, with kinetic freezing often viewed as a less severe instance compared to akinetic. Therefore, assessing the degree of each FoG subtype may help to align an automatic severity assessment more closely with current clinical methods [16, 17]. Additionally, the exploration of ternary FoG may facilitate reflection on both the current clinical definitions of freezing and the performance of models in current literature in the ternary domain (i.e., identifying if existing architectures and feature sets are preferable for specific FoG subtypes) [13, 27].

1.3 Wearable Sensors

Wearable sensors are compact electronic devices equipped with various sensors that can be easily worn or attached to the body. These devices are designed to collect and monitor a wide range of physiological data, such as heart rate, body temperature, movement patterns, and even brain activity. Wearable sensors have gained immense popularity in recent years due to their potential to revolutionize healthcare by providing continuous and non-invasive monitoring of an individual's health and well-being.

In the context of aiding PD assessment, wearable sensors offer a promising avenue for early diagnosis and continuous monitoring of patients. Symptoms from PD can potentially be characterized by subtle changes in movement patterns, cognitive abilities, and even physiological responses. Wearable sensors, such as smartwatches and activity monitors with built-in accelerometers and gyroscopes, can track changes in gait and tremors, enabling early detection or severity tracking throughout the disease progression. By providing real-time data and long-term insights, wearable sensors offer valuable information to clinicians, enabling them to personalize treatment plans and improve the quality of life for patients suffering from these challenging conditions. As technology advances, wearable sensors hold the potential to play a pivotal role in early detection, continuous monitoring, and management of neurodegenerative diseases.

1.4 Classification

Classification is a technique in machine learning involving grouping distinct samples into defined groups. It is a widely used technique in various fields to organize and categorize data based on specific characteristics or features. To achieve classification, a subset of labelled examples, often referred to as the training data, is used in conjunction with a parametric or non-parametric model. This model is designed to learn and capture patterns from the input data which permits differentiation between unique classes. Once trained, the classification model can automatically detect the class of a new sample based on its defining features [4].

One practical application of classification models is in conjunction with wearable sensors. These compact electronic devices equipped with various sensors can continuously monitor physiological data, movement patterns, and other relevant information from individuals. By leveraging classification algorithms, the collected data can be analyzed and categorized to aid in various healthcare applications. For instance, wearable sensors integrated with a classification model can help in the detection and continuous monitoring of FoG symptoms for the PD population [26].

1.5 Thesis Objectives

Based on the current state of FoG classification, the objective of this work is to investigate the ternary domain FoG classification with wearable sensors through current classical and deep machine learning architectures. This investigation contains the following sub-objectives: 1) identification of suitable FoG datasets that can be used for training and testing of machine learning models, 2) selection of both classical and deep machine learning models based on binary classification in current literature, and optimizing the hyperparameters based on the selected datasets, 3) evaluation of model performance on a holdout set of participants, and the impact motor situation and subtype of the freezer have on the performance, and finally 4) the identification of potential applications of the ternary models based on the observed performance.

An additional side-objective was made to explore the impact of solely using a lumbar triaxial accelerometer has on model performance. While this was not the main objective of this thesis, this analysis was included to provide preliminary insight into the tradeoff between the size of the sensor set used during collection and the potential impact on classification performance.

Chapter 2

Literature Review

2.1 Freezing of Gait

FoG in PD is defined as a "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk" [13, 14]. Some individuals describe this phenomenon as the feeling of your feet being glued to the floor despite the intention to make your next step. These events can be categorized into two main subtypes: 1) akinetic, where there is a complete absence of movement, and 2) kinetic, which involves a lack of "effective" forward progression with lower limb movements such as tremors [13]. However, due to considerable heterogeneity within the FoG population, there is no concrete definition for distinguishing the onset, offset, and subtype of freezes, leading to variations in professional interpretation [13]. For example, some may consider akinetic freezes as no movement whatsoever, whereas others may still classify a freezing bout as akinetic if there are small sections of movement within a freezing event.

Moreover, the triggers that provoke FoG events are often unique to each individual, encompassing factors such as the environment, motor context, and emotional state [28, 14]. Researchers have identified three main subgroups based on a set of an individual's triggers, namely asymmetric-motor, anxious, and sensory-attention, which can aid in further characterizing a freezing situation. The derivation of these subgroup labels was accomplished through an individual's C-FOG questionnaire answers. The asymmetric-motor participants represented a greater proportion with asymmetric impairments rather than bilaterally, the anxious group represented individuals with greater scores for anxious-related items, lastly, the sensory-attention group represented individuals with greater scores for set-shifting related items [22].

Both the FoG subtypes and the specific provoking triggers can be correlated with the severity of the symptom. Consequently, understanding these relationships can lead to an enhanced standard of care by optimizing interventions, rehabilitation therapy, and medication strategies to better address the unique challenges presented by FoG in PD.

2.1.1 Clinical Assessment

Currently, the assessment of FoG severity in a clinical setting can vary among clinicians, with no established gold-standard method [19]. Typically, severity is evaluated using questionnaires, such as the NFOG questionnaire, and/or by measuring the percentage of time an individual spends in specific FoG subtypes and FoG onsets and offsets of specific freezing subtypes during tasks designed to provoke FoG events, such as the Ziegler protocol [19, 20, 17, 10]. These tasks simulate common activities of daily living while eliciting FoG events. In clinical research studies, video data from these tasks is often used to analyze an individual's gait patterns, allowing for a more detailed assessment of severity [29]. However, during regular clinic visits, video data collection is resource-intensive and may not be feasible in real time.

Translating the onsets and offsets of each specific freezing subtype into severity is done in various ways [16, 10, 17, 18]. One such method is the Ziegler severity score which looks at the freezing subtypes that occurred and their frequency/duration towards calculating a score out of 36. The hierarchy of the freezing is as follows: 1) no FoG or festination, 2) kinetic FoG - small disturbed and rapid steps, with some degree of advancement, 3) stationary FoG - either trembling in place or total akinesia that patient could overcome, and 4) abortion of the task or interference by examiner. More details around the specific scoring can be found in the work done by Ziegler et al., and Goh et al. [16, 10].

Another method of FoG severity characterization is to bin the time spent frozen for each subtype into distinct groups using threshold techniques. One interpretation of these bins, known as the Jerusalem protocol, breaks up the time spent frozen with less than 10% indicating mild severity, between 10% and 50% indicating moderate severity, and anything above 50% indicating severe impairment. This interpretation is currently in the validation stage of development and has received funding from the Michael J. Fox Foundation to generate evidence to support clinical use.

While the current questionnaires for clinical assessment of FoG are useful for identifying FoG type and frequency, they provide limited insight into the progression of the symptom. Additionally, questionnaires can introduce subjectivity, leading to significant variation between individuals [30]. As for in-clinic visits to measure severity, these might not fully capture the true severity of FoG due to limited testing windows and potential variability between in-clinic freezing and at-home freezing (e.g., Hawthorne effect), potentially creating bias in assessing an individual's true severity [31]. Hence, developing low-cost, low-burden methods of objectively estimating FoG severity is a key objective in PD research.

2.1.2 Remote Monitoring

An alternative to in-clinic assessment is remote monitoring, which can supplement the ability to detect changes in FoG severity and reduce bias from actual FoG severity during a clinical visit. Remote monitoring can be achieved by equipping individuals with a sensor set, such as an inertial measurement unit (IMU) and/or electrocardiogram (ECG), or by capturing video footage during their daily activities. However, remote monitoring is challenged by high dependence on activities performed in daily life with little structure compared to in-clinic protocols. The lack of consistent self-selected activities can lead to inconsistencies in interpreting wearable sensor data and may bias assessments.

Analyzing longer free-living datasets for remote monitoring can be more tedious compared to small in-clinic trials, making it challenging to accurately assess an individual's condition. As a result, the progression of remote monitoring has been limited, with most data collected in remote settings used primarily for exploratory studies. For example, Mancini et al. conducted a study to investigate gait features such as the number of turns and gait bouts, gait speed, and turn angle over a 7-day span. They found correlations between these gait features and the predicted time spent frozen by their binary FoG classification method [32].

Since remote monitoring is often limited to biosignal or IMU data, with motion capture and video data being largely infeasible in free-living scenarios, remote classification mainly focuses on capturing gait analytics [33]. These gait analytics include rudimentary features such as walking speed, step times, or turning speed as proxy measures for FoG severity. As the number of analytics extracted grows, the manual interpretation of the metrics becomes more difficult, especially for longer collection periods. Therefore, approaches involving automatic classification models that utilize these metrics to predict actual FoG events have been investigated, which can then estimate severity through the percentage of time spent frozen [26, 34, 32]. Utilizing a classification model in remote monitoring can help ensure an individual's severity is captured with limited required manual intervention. Not only would such a model provide severity estimations in remote settings where manual annotations would be tedious, but could also serve to further inform clinical assessments.

2.2 Freezing of Gait Classification

In the existing literature, the majority of FoG detection models focus on binary detection [26]. This approach involves consolidating all subtypes of FoG into a singular term, resulting in only two labels: no FoG vs FoG. While this simplifies the classification problem, it comes at the expense of losing additional insight that could be gained from identifying types of FoG. In particular, separating the FoG subtypes will help to more closely align remote monitoring with clinical assessments that weigh subtypes differently when calculating the severity [16, 17].

2.2.1 Data Selection

The performance of a machine learning model is highly influenced by the quality of data; therefore, in FoG classification, the selection of high-quality datasets is imperative for efficient and accurate predictions. The data selection process consists of two main parts: 1) the sensor type and locations used to collect the raw data, and 2) pre-processing techniques applied to the raw data before passing it to the model.

In current literature, the most frequent sensor set used is a tri-axial lumbar accelerometer, likely due to low obtrusiveness, low burden to the individual and low cost of equipment. Other sensors have been explored, such as ECG, IMU with gyroscopes, or multiple sensors on other locations (e.g., bi-lateral shank or feet [26]). The sensor setup utilized by a model can have a large impact on the performance. For example, a strong sensor set should capture descriptive features that have a high correlation to FoG events. However, there is a trade-off: a lower number of sensors brings lower cost and high ease of use for the PD individual, whereas a larger sensor set can provide more descriptive feature sets and more accurate models. It should also be noted that the impact the size of the sensor set will have on the performance will in fact have a saturation point (e.g., adding another IMU near an existing sensor, such as an additional foot sensor, will add little to no additional information).

For pre-processing techniques, methods used in the literature are dependent on the type of model used. Many efforts have made use of filtering, normalization, and/or standardization techniques to help remove noise and ensure all signals are equally scaled. Sensor signals are often segmented into windows of time to allow for feature calculations and/or models that require inputs of more than a single sample instance. The size of these windows heavily depends on the model architecture and features planned to be extracted. For FoG classification, previous works have utilized window sizes between 2 and 4 seconds of data [26, 35, 36].

Some models, mainly classical types, require features to be passed as inputs instead of the original raw signal. Unlike deep learning models, classical models lack the ability to automatically extract (or learn) information from raw signals. The extraction and selection of the parameters prior to the model is the process of feature engineering, which aims to extract useful information from each subsequent window.

Some of the most common features in current literature are shown in Table 2.1. All features from this table were based on the popularity observed in the literature review work done by Pardoel et al., and the FoG classification work done by Arami et al., with only applicable sensor locations and sensor types relative to the accessible datasets of this study included [26, 34]. This inclusion criterion for features selected consists of accelerometer and gyroscope sensors, located on bilateral feet, lumbar, and/or sternum. In combination with these criteria, features requiring a separate classification algorithm, such as gait detection and analytics pipeline, were left out of the table due to the possibility of compounding errors.

In combination with the features extracted, some studies have utilized correlation statistics to identify an optimal feature set for FoG classification [34]. This feature pruning process can be broken into 3 categories including filters, wrappers, and embedded techniques. For purposes of this work, only the filter methods are explored. Filter pruning involves calculating scores for each feature and selecting only the top k-performing feature set [37]. This score can be calculated in a univariate sense, where each feature vector is compared to the true labels. From here, statistical methods such as Kendall's or Spearman's correlation, analysis of variance (ANOVA), or mutual information can be used to calculate a score [38, 39, 40]. Another class of methods within the filter technique is multivariate approaches, which also compares each feature to the true labels, but also uses the relationships between features to generate the score. While this process is more involved and computationally expensive, it can identify a more optimal set of features. An example

Feature	**Sensor Type**	**Description**
Mean	Acc., Gyro.	Mean of signal
Min, Max, Median	Acc.	Descriptive statistics
RMS	Acc., Gyro.	Root mean square (RMS) of a signal
Standard Deviation/Variance	Acc., Gyro.	Standard deviation of a signal
Integral	Acc.	Integral of acceleration (singular or double)
Kurtosis	Acc., Gyro.	Measure of signal tailedness within window distribution
Skewness	Acc.	Measure of signal asymmetry within window distribution
K index	Gyro.	Summation of the absolute value of low pass filtered angular velocity of left and right shanks in sagittal plane
Number of dominant peaks	Acc.	Number of signal dominant peaks
Number of zero crossings	Acc.	Number of signal zero crossings
Pearson's correlation coefficient	Acc., Gyro.	Similarity between two signals
Entropy	Acc., Gyro.	Shannon's entropy calculation
PSD bands	Acc.	Specific frequency bands of power spectral distribution (PSD)
Band Power	Acc., Gyro.	Area under the curve of power spectral density plot, between specific bands
Freezing Index	Acc.	Ratio of signal power in freeze band (3–8 Hz) and locomotion band (0–3 Hz)
Extended Freezing Index	Acc.	Square of the freezing index with a prior mean subtraction

Table 2.1: Common features for accelerometer (acc.) and gyroscope (gyro.) sensors in current literature.

of this technique is called minimum Redundancy - Maximum Relevance (mRMR), which aims to identify a multi-variate correlation between features and the labels to identify the most descriptive and the least redundant feature set [41]. As for the models not requiring pre-processed features, such as deep models, samples are often similarly windowed and have pre-processing techniques such as filtering or normalization applied to improve the learning capabilities of the model.

The last component of the data selection process is annotating or labelling of the data. High-quality annotations are imperative for supervised learning techniques discussed in this work and, therefore, should not be overlooked when designing the classification pipeline. Often, annotations for FoG classification are carried out in a similar manner to clinical studies, where multiple raters use video data to flag all FoG events during a task with a clear process to handle non-consensus labels [42, 29].

2.2.2 Binary Models

The majority of existing FoG classification models focus on the binary paradigm by combining all FoG subtypes into a single label. These binary models can be further broken down into three main categories: 1) threshold, 2) classical, and 3) deep models. For threshold methods, there is no required training set and the models are designed around fixed parameters prior to implementation. In this work, threshold methods were not explored due to being out of the scope of investigating machine learning models for FoG detection. As for classical models, these can be parametric or non-parametric, but regardless require a distinct set of samples for making inferences on new samples. The term classical describes the type of architectures used within this category, which are more traditional mathematical algorithms such as logistic regression (LR), random forest (RF), or support vector machine (SVM), along with the basic variation of a neural network (NN) with the limitation of at most one hidden layer. Finally, deep models are similar to classical models with the requirement of a set of samples in order to train a model. The only deviation is the complexity of the architecture, with deep models utilizing networks with two or more hidden layers which help to establish deeper connections between the internal nodes.

2.2.2.1 Classical Models

As stated, classical models include parametric or non-parametric models utilizing a training set of samples to make inferences on new samples with traditional mathematical algorithms, or simple (i.e., few hidden layers) feed-forward NN architectures. Classical models offer the benefit of relative simplicity, which not only helps reduce the amount of overfitting but also lowers the conceptual complexity, making them advantageous in medical applications like FoG classification. This simplicity also facilitates the development and training of models on limited datasets, a common scenario in medical applications.

Several classical models are commonly used in FoG classification, including LR, RF, SVM, AdaBoosted decision tree (ADT), and k-nearest neighbours (KNN) [26]. However, the reported performance of these models varies widely in the literature due to differences in study designs, definitions for annotating FoG events, and the quantity of data available. Decision tree-based models, such as RF and ADT, have demonstrated sensitivity performance ranging from 66.25% to 98.35% and specificity ranging from 66.00% to 99.72%. On the other hand, SVM models have achieved sensitivity ranging from 74.7% to 99.73% and specificity ranging from 79.0% to 100% [26]. These performance ranges were determined through various testing approaches, including cross-validation, leave one subject out (LOSO), and test hold-out sets [26].

2.2.2.2 Deep Models

Deep models are similar to classical NN models, however, they involve numerous hidden layers and are often equipped with techniques such as convolutional layers, skip connections, or recurrent layers, instead of the basic feed-forward structure. The benefits of deep models are on the opposite side of the spectrum compared to classical models with high complexity due to the increase in the number of layers and connections. This increase in complexity, in turn, helps to reduce model bias but increases the variance of overfitting. To help limit undesired variance, larger datasets with high variability are required; additionally, techniques such as early stopping during training can be implemented [43, 44, 45, 8].

As mentioned earlier, another benefit of deeper models is the ability to learn features internally, requiring only the windowed signals as inputs to the model. Deep models that utilize convolutional layers in their backbone are capable of this feature extraction; however, a convolutional backbone can be inserted prior to most architectures to replace the external feature engineering step.

Additionally, one concept that is included in specific types of deep models is the ability to look at multiple windows at once. Considering multiple windows at once in a sequence format is not limited to deep models. Features from multiple windows can be passed into classical models, although, this is not as common as the deep network architectures. The concept of passing multiple window samples for a singular label introduced through recurrent neural network (RNN) has the benefit of utilizing neighbouring temporal information while keeping the window sizes in the desired range [46]. This method is preferred since increasing window size reduces the temporal resolution of the classification pipeline.

Some of the common deep models in the literature include convolutional neural network (CNN) and transformers [43, 44, 45, 8]. As opposed to classical models, similar model architectures are still unique due to the complexity and variations in the design process. Therefore, instead of reporting ranges, individual performance found for various deep network models in literature can be seen below in Table 2.2.

Model Type	**Group**	**Sensitivity**	**Specificity**	**Precision**	**F1-Score**
CNN	Borzi et al.	0.877	0.883	-	-
CNN + MLP	Bikia et al.	0.856	0.857	-	-
CNN	Yang et al.	-	-	-	0.67
Transformer	Sigcha et al.	0.842	0.939	0.617	0.71

Table 2.2: The performance of recent deep model architectures for binary FoG classification.

2.2.3 Ternary Models

At the time of writing, all FoG models in literature have been developed for the binary case where all subtypes are consolidated into a singular FoG label. This gap in the field warrants the exploration of the performance of successful models in the binary space extended for the ternary space. By characterizing the performance of extended models, the findings of the current study will help provide insight into their capacity to detect and classify FoG.

Ternary classification is important for many reasons, primarily to provide greater insight to clinicians. This additional insight is mainly in the form of a more granular assessment of an individual's disease progression by providing labels and severity estimations for each FoG subtype. Pure akinesia is often viewed as the most severe form of freezing, therefore combining the kinetic and akinetic severities into a single metric may be misleading from an individual's true severity and quality of life [16, 17]. Thus, this additional insight can help better identify and plan specific treatment(s) or disease management plans (e.g., medication, therapy) to suit an individual's situation.

Lastly, as mentioned due to the large heterogeneity in the FoG population, the exact definition for the subtype, onset and offset of FoG have various interpretations [13]. Therefore, the exploration of ternary FoG classification also aims to advance the understanding and relationships between the subtypes by analyzing how trained models from the binary space perform in the ternary domain.

2.2.4 Motor Context

Transitioning away from solely defining the problem of FoG as ternary or beyond, it is crucial to explore another aspect often overlooked in current literature: the performance of FoG classification in relation to the motor context. While some studies, such as the work by Yang et al., test their models on various gait tasks like the timed up and go (TUG) and turning separately, this limited investigation merely scratches the surface of understanding the impact of the motor context [44].

Understanding the implications of the motor context on classifier performance is particularly significant when considering the application of FoG detection. In particular, if deployed in uncontrolled free-living environments, it is important to understand in what motor situations (e.g., turning, straight-line walking, approaching a doorway) a model will succeed, and in which situations it has trouble. Not only is this applicable for free-living conditions, but as well, severity estimations are often broken down by motor situations to gain a better understanding of an individual's triggers [18, 22]. With the model's motor situation performance, the shortcomings of the model in a motor situation can potentially be addressed, or be used to develop a system that utilizes the best model for each specific motor situation.

Chapter 3

Data

3.1 Sydney Dataset

The first dataset used in this work was collected by Goh et al. at the University of Sydney in Australia. The collection of this data was completed with the intention of assessing the reliability of the protocol for evaluating FoG in people with PD. It also aimed to determine correlations between the test results and video annotations, examine the test duration as a measure of FoG severity, assess usability, and explore reliability differences based on clinical experience. This dataset consisted of 10 participants, all deemed freezers in the PD population based on their clinical history and questionnaires (e.g., Movement Disorder Society-Unified (MDS-Unified) Section III, and New Freezing of Gait Questionnaire) [10].

3.1.1 Protocol

Each participant who participated in the dataset's study completed the Ziegler protocol in both their "ON" medication and "OFF" medication states [16]. The Ziegler protocol involves the following steps: 1) starting the trial seated, 2) after the trial is started, the participant stands from their seat, 3) once standing, they walk forward until they are within a box taped/painted on the ground, 4) while within the box, they complete one 360 degrees turn clockwise, then one 360 degrees turn counterclockwise, 5) after turning, they continue to walk forward until approaching a closed door, 6) once at the door, they open it, walk through the doorway, and perform a 180-degree turn, and 7) once all of this is complete, they make their way back to their seat, and the test concludes once they pass the box where the 360-degree turns were made [16, 10, 1] (see Figure 3.1).

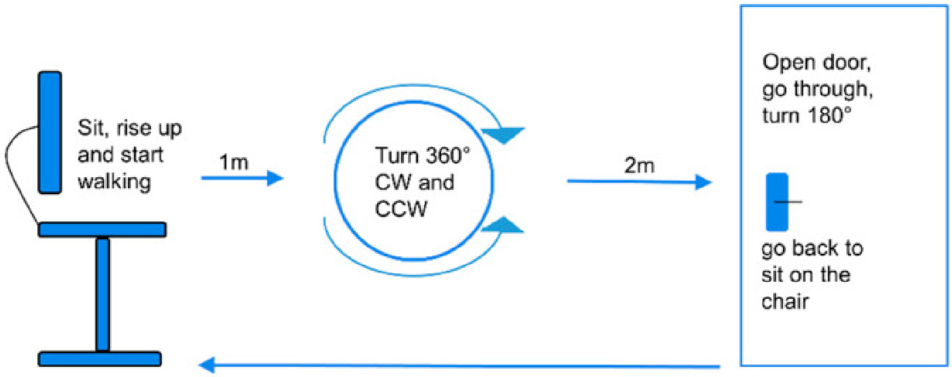


Figure 3.1: Illustration of the tasks completed in the Ziegler protocol [1].

The Ziegler protocol was completed three times for each medication state, with each iteration adding an additional cognitive task to increase the probability of provoking a FoG event. The first iteration was the regular Ziegler protocol with no additional cognitive loads. The second iteration involved the addition of carrying a tray while completing the Ziegler protocol. The final iteration kept the tray while adding the requirement of having the participant complete a counting task during the trial. Not every participant completed all three variations in the "OFF" medication state due to increased amounts of freezing. Individuals who spent the majority of their regular Ziegler trial in the "OFF" medication state were not required to complete the subsequently increased cognitive load iterations. This occurred for two participants, resulting in a total of 54 trials being collected.

3.1.2 Hardware

During the trials, the participant was wearing seven Opal IMU sensors located on the forehead, sternum, lumbar, and bilateral shank and foot. Each IMU sensor captured tri-axial acceleration and angular velocity at 128Hz. The bilateral shank IMU data was excluded from this work due to some a subset of trails missing data. Additionally, video data was captured for the purposes of annotating each trial to generate FoG, and motor situation labels.

3.1.3 Annotations

FoG annotations were made by multiple FoG experts by breaking each video into 3 checks including 1) identifying each situation throughout the trial (i.e., no FoG, festination, FoG and the specific subtype, and deviation from the protocol), 2) The start and stop of each situation, and 3) any additional comments [10, 29]. With this structure, each trial was annotated for the start and stop of FoG events and their respective subtype. Lastly, the annotators produced a Ziegler severity score out of 36 for each individual. To ensure consistency in the rating, the interrater reliability was determined using intraclass correlation coefficients and was found to be 0.8. More details on the annotation procedure can be found in the supplementary materials for the work done by Goh et al. “The Ziegler Test Is Reliable and Valid for Measuring Freezing of Gait in People With Parkinson Disease” [10].

As well, the motor situation during the trial was annotated for future performance breakdown based on the context. These annotations were made outside the Goh et al. group by members of the Neural and Rehabilitation Engineering (NRE) lab at the University of Waterloo, applying a similar annotation breakdown structure to the FoG labels [10, 29]. To limit the variability, each trial was broken down into four distinct sections including 1) forward walking from the first effective step after standing to reach the box where the 360-degree turn is completed, 2) right turn from when they initiate the clockwise turn to completion, 3) left turn for when they imitate the counter-clockwise turn to when they make a step out of the box, 4) doorway for the remainder of the trial. These 4 labels (forward walking, right turn, left turn, and doorway) span the entire trial with no gaps between neighbouring labels, except for the initial sit-to-stand prior to forward walking.

3.1.4 Descriptive Statistics

Throughout the entire Sydney dataset, a total of 24.9% was spent in an akinetic freeze and 8.87% was spent in a kinetic freeze. The breakdown by participant can be seen in Table 3.1. Additional participant statistics can be found in Table 3.2. For more demographic statistics, please refer to the work done by Goh et al. "The Ziegler Test Is Reliable and Valid for Measuring Freezing of Gait in People With Parkinson Disease" [10].

Participant Code	Akinetic Severity	Kinetic Severity	Number of Trials
21DH	Mild (0.71%)	Mild (10.0%)	6
28FV	Moderate (32.5%)	Moderate (16.92%)	6
39KR	Severe (76.55%)	Mild (1.35%)	4
45PG	Mild (0.0%)	Moderate (20.14%)	6
54EJ	Mild (2.29%)	Mild (9.17%)	6
76CA	Mild (2.49%)	Mild (3.56%)	6
83OS	Mild (1.01%)	Moderate (16.58%)	6
85TL	Mild (0.0%)	Mild (0.76%)	6
93QN	Mild (0.0%)	Mild (0.0%)	6
97MU	Severe (64.57%)	Mild (2.29%)	4

Table 3.1: Breakdown of the severity and number of trials for all individuals within the Sydney dataset.

Participant	Sex	Age	Duration of PD (years)	Most affected side	C-FOG Dominant Subgroup	UPDRS (0-132)	MMSE (0-30)
21DH	Male	78	11	Left	Sensory Attention	29	30
28FV	Male	81	12	Both	Asymmetric Motor	53	26
39KR	Female	75	8	Left	Sensory Attention	27	28
45PG	Male	62	14	Left	Anxiety	18	30
54EJ	Male	66	16	Left	Anxiety	35	25
76CA	Male	75	16	Both	Sensory Attention	57	28
83OS	Male	76	5	Left	Anxiety	45	29
85TL	Male	60	10	Right	Sensory Attention	47	29
93QN	Male	61	18	Right	Sensory Attention	40	30
97MU	Male	72	23	Both	Asymmetric Motor	22	28

Table 3.2: Participant demographics, and PD-related information for all individuals within the Sydney dataset. C-FOG dominant subgroup was based on the scores on page 2 out of 4 for the C-FOG questionnaire [10].

3.2 MJFF Dataset

The second dataset utilized in this study, known as the MJFF dataset, comprises multiple collections from various groups [47]. These groups include The Center for the Study of Movement, Cognition, and Mobility, The Neurorehabilitation Research Group at Katholieke Universiteit Leuven in Belgium, and the Mobility and Falls Translational Research Center at the Hinda and Arthur Marcus Institute for Aging, affiliated with Harvard Medical School in Boston. With the assistance of the Michael J. Fox Foundation for Parkinson's Research, the data from these sources were merged and made available on a Kaggle competition for developing FoG algorithms.

The MJFF dataset consists of 924 trials completed by 100 participants. This dataset plays a crucial role in the study by expanding the number of trials from the original 54 to close to 1000 trials total when combining the datasets. This is imperative since this extensive dataset allows for the training of deeper models with large dataset requirements.

3.2.1 Protocol

The data for this study was collected from multiple sources, the dataset was divided into two sub-datasets with different protocols. The first sub-dataset, named tDCS, followed the Ziegler protocol and was conducted in both the "ON" and "OFF" states with three levels of difficulty, similar to the Sydney dataset [1, 48]. The second sub-dataset, named DeFOG, employed a different protocol designed to provoke freezing of gait (FoG) events through various gait tasks, also in both the "ON" and "OFF" medication states [2, 47]. The DeFOG protocol consisted of seven specific gait tasks completed by participants in their own homes. These tasks included 4-meter walk, Timed Up and Go (TUG), dual-task TUG (involving subtracting numbers as an additional cognitive load), turning tasks with alternating directions, dual-task turning tasks with cognitive load, hotspot door (a walking trial involving opening a door, entering another room, turning, and returning to the start point), and personalized hotspot (walking through an area in the house that the subject described as FoG provoking). For the second visit, the same tasks were completed for both medication states, but with the inclusion of audio cueing for the non-dual-task tasks. The audio cueing involved providing tones at certain frequencies when a FoG event was detected to help re-initiate walking [47]. A graphic illustrating this DeFOG protocol is presented in Figure 3.2.

The number of trials completed for each protocol is detailed in Table 3.3.

3.2.2 Hardware

For each protocol completed in this dataset, a different hardware set was used, however, all data collected in the MJFF dataset only used a single lumbar tri-axial accelerometer sensor. For the Ziegler protocol, an Opal tri-axial accelerometer was used at a sampling rate of 128Hz. As for the DeFOG protocol, this data was collected with a tri-axial Ax3 by Axivity at 100Hz. Similar to the Sydney dataset, video data was used for later offline annotations of each trial [47].

3.2.3 Annotations

Annotations were made through offline video analysis of each trial. The annotations included the start and stop of the FoG event, and whether or not it was kinetic or akinetic.

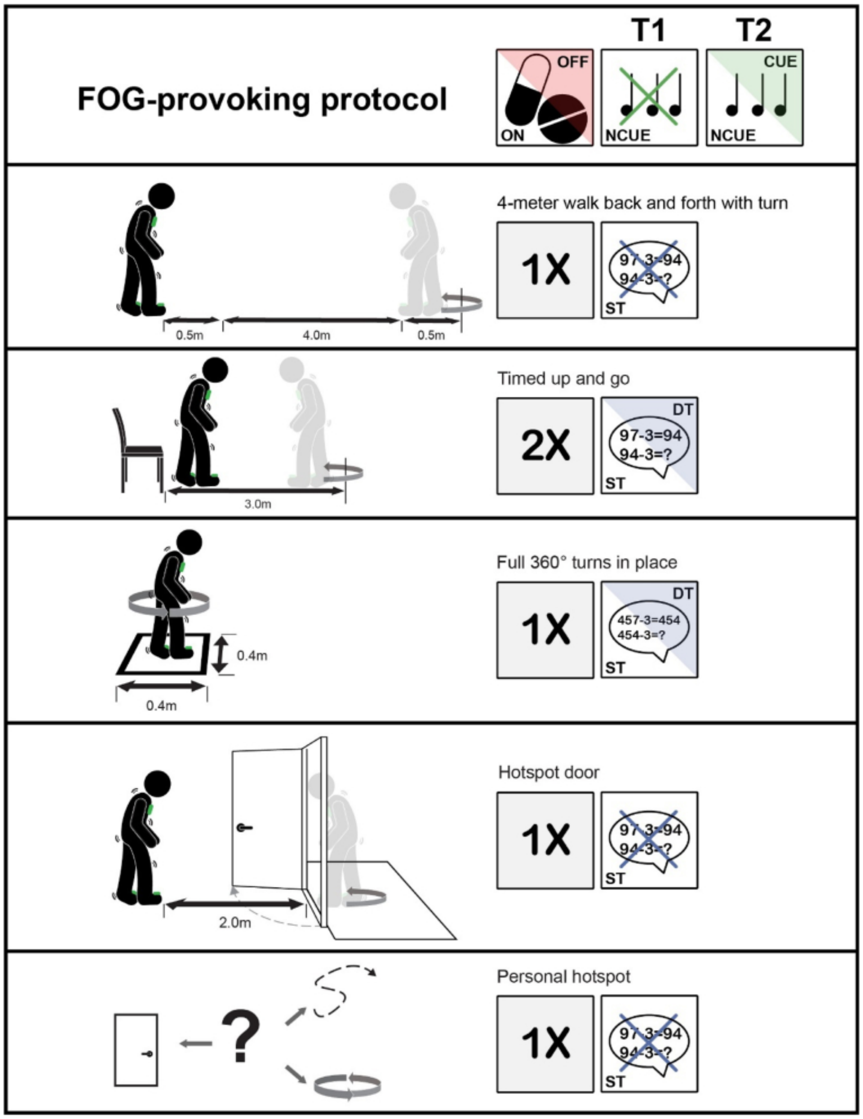


Figure 3.2: Illustration of the tasks completed in the DeFOG protocol [2].

No labels were made to highlight the motor situation of the individual throughout the trial. However, additional annotations were made to highlight where each freeze occurred within the options of start hesitation, turning, or walking [47]. These labels could have been used to produce a general understanding of the motor situation, however, this was not pursued due to the labels only being present during FoG events which would cause the label quantity per trial to be inconsistent.

3.2.4 Descriptive Statistics

Throughout the entire MJFF dataset, a total of 1.47% was spent in an akinetic freeze and 12.39% was spent in a kinetic freeze. The breakdown of akinetic and kinetic freezes within each sub-dataset, along with the percent each sub-dataset contributes to the overall MJFF dataset can be seen in Table 3.3. Additionally, the distribution of severities for akinetic and kinetic across all participants in both protocols can be seen in Figure 3.3. This distribution binned the percent time spent frozen for akinetic and kinetic events separately for each individual based on the Jerusalem protocol mentioned earlier. For the participants of the MJFF dataset, the average age was 67.8 ± 7.95, and the average years since diagnosis was 10.4 ± 6.27.

Dataset	**Akinetic Percent**	**Kinetic Percent**	**Number of Trials**
tDCS	0.73 %	30.30 %	833
DeFOG	1.85 %	3.04 %	91

Table 3.3: Breakdown of the freezing distribution between the two sub-datasets within the MJFF dataset.

3.3 Preprocessing

During the preprocessing phase of the data, specific steps were applied to the datasets to ensure proper synchronization and prepare them for further analysis [8, 49, 50]. For the Sydney dataset, synchronization was achieved by identifying an auditory tone in the video data, which marked the beginning of each trial. This tone was used to crop both the annotations and sensor data accordingly. As for the MJFF dataset, synchronization had already been carried out by external sources, eliminating the need for this step.

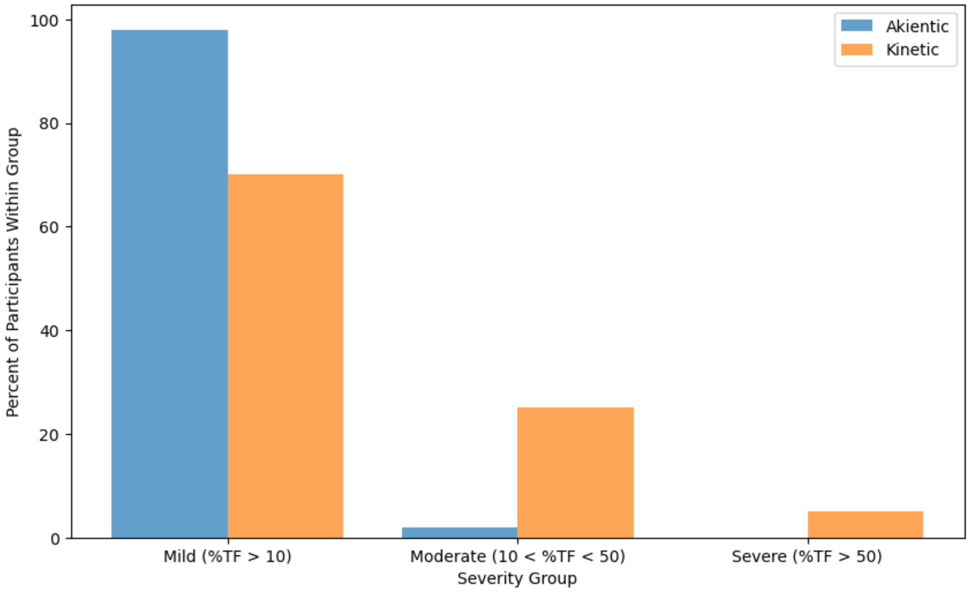


Figure 3.3: Distribution of the participant severities for akinetic and kinetic freezing in the MJFF dataset based on the total time spent frozen (%TF).

After synchronization, the raw data and labels for each dataset were subsampled to a rate of 40Hz [8, 49, 50]. This subsampling ensured computational efficiency while still preserving the essential frequencies for human activity recognition and analysis, particularly in the freezing of gait (FoG) domain, which involves frequencies in the 3-8 Hz range. To facilitate consistent comparisons, raw data provided by the IMU sensors was normalized within the range of [0,1] with respect to each individual trial, minimizing variability introduced by different sensor types.

The next step involved windowing the data to create chunks of samples with a window size of 3.2 seconds and a step size of 1.6 seconds [26, 36, 35, 8]. This window size was selected based on the existing literature, falling within the range of 2-4 seconds. Moreover, using a window size corresponding to an integer power of 2 (e.g., 3.2 seconds at 40 Hz gives 128 samples) allowed for efficient fast Fourier transform (FFT) calculations in subsequent analyses.

Similarly, the labels were windowed with the same specifications: 3.2 seconds window size and 1.6 seconds step size [26, 36, 35, 8]. For the windowed labels, a single label was generated by taking the mode of the window, representing the label that appeared most frequently in a given window. Additionally, it was noted whether this label corresponded to the second half of the window (i.e., the last 1.6 seconds). This approach aimed to reduce overlap during the unwindowing process in postprocessing. An issue arising from this method was the vacant label period during the first half of the first window in each trial. To address this, each label array was padded with non-FoG labels, indicating a lack of freezing, thereby ensuring consistency in subsequent analyses.

3.3.1 Classical Model Preprocessing

The preprocessing applied to both datasets that is specific to the classical models was the feature extraction. This process involved extracting a feature set from the windowed data, which could then be passed into the classical models for training and evaluation. More details on this process can be found in the subsequent section 3.4.

3.3.2 Deep Model Preprocessing

As mentioned earlier, the deeper models utilized convolutional layers in the network backbone to automatically extract relevant features, eliminating the need for external feature engineering. Consequently, the preprocessing steps for the deeper models diverged from those of the classical models.

Specifically, for the deeper model, the data underwent a transformation into the frequency domain using the FFT. This choice was influenced by the findings of Sigcha et al., who demonstrated that representing data in the frequency domain led to superior performance compared to using time-domain signals and features [51].

Once the data was transformed into the frequency domain, the windows were grouped into sequences of four consecutive windows. The labels for these sequences were derived from the last window in each series. However, this approach introduced a gap in the labels at the beginning of each trial, as the first three windows in each sequence had no associated inferences. To address this, we padded the initial sequences with zeros, indicating the default non-FoG label.

For more detailed information on the architecture and design of the deeper model used in this work, please refer to section 4.2.

3.4 Feature Engineering

The feature engineering process commenced by identifying a large pool of potential features (see Table 3.4). The pool consisted of 24 features when considering the tri-axial lumbar accelerometer alone and expanded to 145 features when incorporating the tri-axial IMU sensor data from the sternum, lumbar, and bilateral feet.

Feature	**Sensor Type**	**Sensor Location**	**Direction**
Root mean squared	Acc., Gyro.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Standard deviation	Acc., Gyro.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Kurtosis	Acc., Gyro.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Skewness	Acc., Gyro.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Number of dominant peaks	Acc.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Number of zero crossings	Acc.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Extended freezing index	Acc.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Displacement	Acc.	Sternum, lumbar, bi-lateral feet	AP, ML, V
Cross-correlation	Gyro.	Bi-lateral feet	ML

Table 3.4: Initial feature pool for accelerometer (acc.) and gyroscope (gyro.) sensors established based on common features in current literature (AP - anterior-posterior, ML - medial-lateral, V - vertical).

Next, the feature set underwent trimming using various methods to determine the optimal selection. For this purpose, the 3-fold cross-validation F1-score was computed for the classical models, employing their default scikit-learn hyperparameters [52]. To simplify the feature engineering process, the individual F1-scores of the multiple classical models were averaged.

The initial evaluation involved the complete pool, establishing a performance benchmark without employing feature selection. Subsequently, feature sets comprising 40%, 60%, and 80% of the features (k-value) were evaluated. This pruning process was performed using five feature selection techniques. The first four techniques, including ANOVA, mutual information, Spearman's, and Kendall's, were univariate approaches relying on the correlation between labels and individual features. The top correlated features were selected and retained for further performance analysis. The fifth technique, mRMR, constituted a multi-variate method, enabling the assessment of correlations not only between features and labels but also within the feature set, thereby identifying redundant features.

The k-value and feature selection method combination that yielded the highest overall cross-validation F1-score defined the final feature set. Additionally, in cases where sensors were symmetrically applied (e.g., bilateral IMU sensors in the Sydney dataset), features extracted from one side were mirrored to mitigate left or right bias in the feature engineering pipeline (i.e., if only the right foot had feature X extracted for the Y direction, then the left foot feature X in the Y direction was added to the feature pool, regardless of the ranking).

For a comprehensive account of the results obtained from this feature engineering study, please refer to section 5.1.1.

Chapter 4

Classification

After collecting data and preprocessing accordingly, the subsequent step in automatic FoG detection is to select and train a model. In this work, six classical machine learning models and one deep network machine learning model were tuned, trained and tested on the data to explore the optimal architecture for FoG detection. The classical models were implemented with the use of the Scikit-Learn framework, whereas the deep network model was implemented with the Pytorch deep learning framework [52, 7].

4.1 Classical Models

As stated, the classical models include architectures that deviate from a neural network (e.g., LR, RF, or SVM), or have less than 2 hidden layers. This makes these models ideal for problems where overfitting is an issue, which can often be present in unbalanced datasets, or smaller datasets. However, the trade reduced overfitting, for often higher bias as a result of the lower complexity. The classical models used in this study pulled inspiration from current literature and include the following: LR, RF, SVM, KNN, ADT and finally, a single layer NN [26, 34].

4.1.1 Logisitc Regression

LR classification builds on the fundamentals of linear regression by using a linear function of the inputs to produce the inference (i.e., $y = wx_i + b$). However, leaving the model with this representation would produce an output outside the classification range (i.e., $(0, 1)$. Therefore, to limit the range, the standard logistic function, otherwise known as the sigmoid function, is used to translate the linear function output within the continuous range to a classification range (see Equation 4.1). One caveat for the sigmoid function is the limitation of a binary output, therefore for classification problems beyond binary one logistic regression model can be created for each class in a one versus rest (OVR) fashion [4].

$$f(y) = \frac{1}{1 + e^{-y}} = \frac{1}{1 + e^{-(wx+b)}} \tag{4.1}$$

4.1.2 Random Forest

A RF classifier in its simplest form is a collection, or forest, of basic decision trees, which is a concept known as ensemble learning. A decision tree thresholds features at each node, and then depending on if the values are higher or lower, the path will be followed. The last layer of nodes is assigned classes, which forces classification by following the directed path along the tree. Along with being a collection of decision trees, a RF classifier uses a method called bagging which only uses subsets of the training data and feature set at each node which helps to introduce variance. The collection of these high variance trees helps to produce an overall model with reduced overfitting after combination with majority voting [4]. An illustration of this ensemble learning technique can be seen in Figure 4.1.

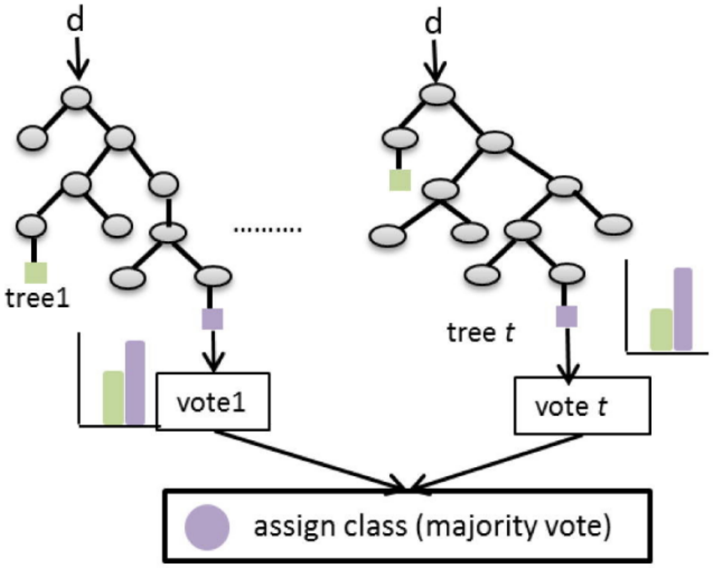


Figure 4.1: Visualization of the random forest architecture, modified from [3].

4.1.3 Support Vector Machine

A SVM classifier functions by representing the n-dimensional feature vector of each sample in an n-dimensional space. From here, a model is developed by identifying a hyperplane that divides the classes best (i.e., the largest margin between the closest opposing class samples). However, sometimes the n-dimensional space is not enough to find a distinguishing hyperplane, therefore SVM models introduce transformations known as kernels. These kernels take the n-dimensional space and translate it into an m-dimensional while introducing non-linearity into the model. Finally, for multiclass problems, this is often treated in a similar way to logistic regression, with solving for the model multiple times in a OVR fashion [4]. A 2-dimension example of a SVM classifier can be seen in Figure 4.2.

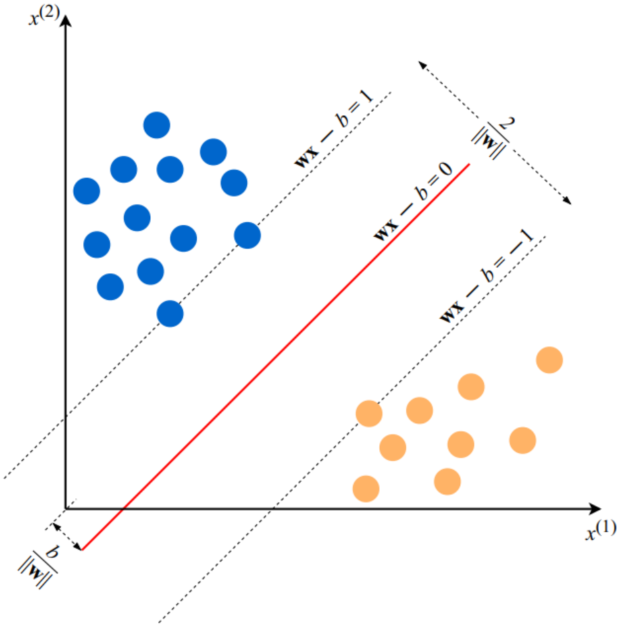


Figure 4.2: 2-Dimensional visualization of a support vector machine model [4].

4.1.4 K-Nearest Neighbnours

KNN differs from the other models considered by being a non-parametric algorithm. KNN represents each sample as a point in n-dimensional space for an n-dimensional feature set. From here, a sample is classified as a certain class based on the proximity to the already defined class distributions. To quantify the proximity, distance functions are used with some of the most common ones being Euclidean distance, and the cosine similarity. The values from the distance function are then used to define a subset of the k-closest labelled points within the space. This subset of k-closest labelled points is then used in a majority voting system to identify which class the sample belongs to. There is a tradeoff when defining the number of neighbouring points to include with higher values being more computationally expensive, but better reflecting the relationship of the sample to the class distribution of the entire population [4]. A 2-dimensional representation of this algorithm can be seen in Figure 4.3.

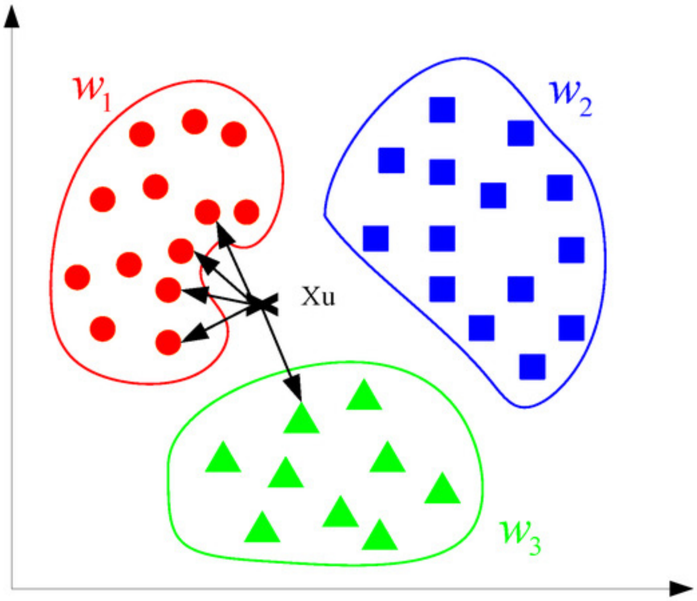


Figure 4.3: 2-Dimensional visualization of a k-nearest neighbours model [5].

4.1.5 Decision Tree with AdaBoost

ADT is very similar to a RF classifier with it being a collection of decision trees combined to produce a single class output. However, the difference for a ADT classifier is often the trees within the collection are extremely simplified (i.e., only one level of nodes). The concept behind this type of ensemble learning is that enough simplified classifiers combined will produce a low variance prediction. In addition to this, ADT also frequently alters the equally distributed majority voting and instead can assign different weights to specific trees within the model. Lastly, unlike a RF model which produces all of the trees in the forest without influence, the previously created trees in AdaBoost impact the development of future trees [53].

4.1.6 Neural Network

The NN classifier is the most complex out of the classical models considered in this work. glsnn function by having a series of nodes at each layer which take in the outputs from the previous layer, transform the values and then pass them on to the next layer. Often each node in layer $i-1$ will pass its output into every node in layer i, this is often called a fully-connected structure (see Figure 4.4). Each node within each layer has weights and a bias and applies it to the input vector in a similar manner to LR (i.e., $y = wx_i + b$). Again, similar to logistic regression, non-linearity is introduced to the system through transformations known as activation functions in the NN space. One of the most common activation functions is ReLU which can be seen in Figure 4.5. To achieve classification, the final layer will utilize an activation function and a threshold so a binary output is produced. For classification problems higher than binary, multiple output nodes can be used, with each representing its own class [4].

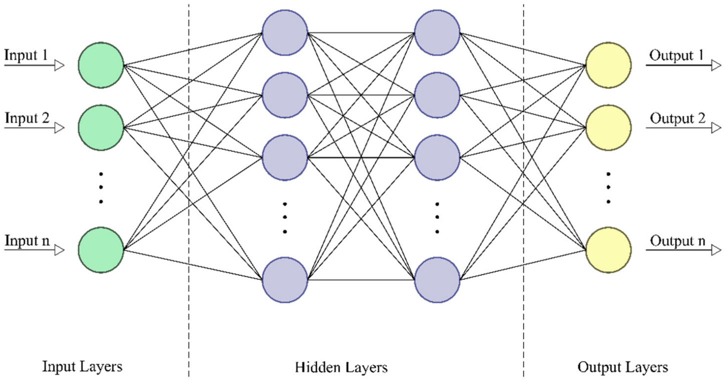


Figure 4.4: Visualization of a fully-connected feed-forward neural network architecture [6].

4.2 Deep Network Model

The term deep network essentially encapsulates any neural network with numerous hidden layers. This produces architectures that often are inversely proportional to classical models,

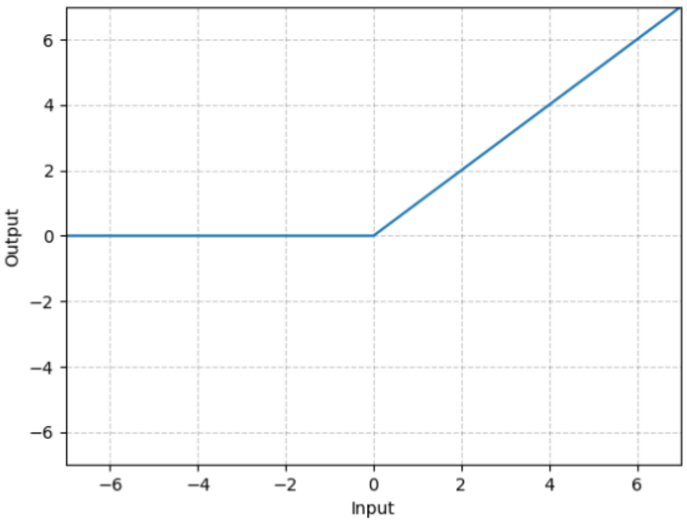


Figure 4.5: ReLU activation function used to introduce non-linearity into neural network models [7].

with a higher chance of overfitting (or variance), and lower bias. However, the increased amount of overfitting can be addressed with larger more variable training datasets, or early stopping while training to prevent this overfit. To implement early stopping, a random stratified 10% of the training data was separated and used to make a validation set. If the validation loss was greater than the training data loss by 0.1 for ten consecutive epochs, then the training procedure was halted. This criterion was defined in a relative sense between the train and validation loss, rather than only considering the loss of the validation because it aimed at limiting the bias towards the training data. Early stopping can also be used with only the validation loss, however, this is for computational resource optimization through stopping training when the loss function appears to be converging to its global minimum.

The deep network model used in this work was modelled after the transformer network developed by Sigcha et al. with slight modifications in the architecture and training hyperparameters [8]. These changes were designed through the tuning procedure, which is explained in section 4.3.2.

4.2.1 Transformer

The transformer developed by Sigcha et al. is a variation of the traditional transformer architecture introduced by Vaswani et al. [54]. In deep models, we can often break the architecture into three sections: the backbone, the neck, and the head [55]. The backbone serves as the feature extractor, while the neck organizes these extracted features, highlighting their similarities and differences. The head then utilizes these organized details for final predictions. For their model, Sigcha et al. used convolutional layers as the backbone, transformer encoder blocks as the neck, and a regular neural network as the head (see Figure 4.6) [8].

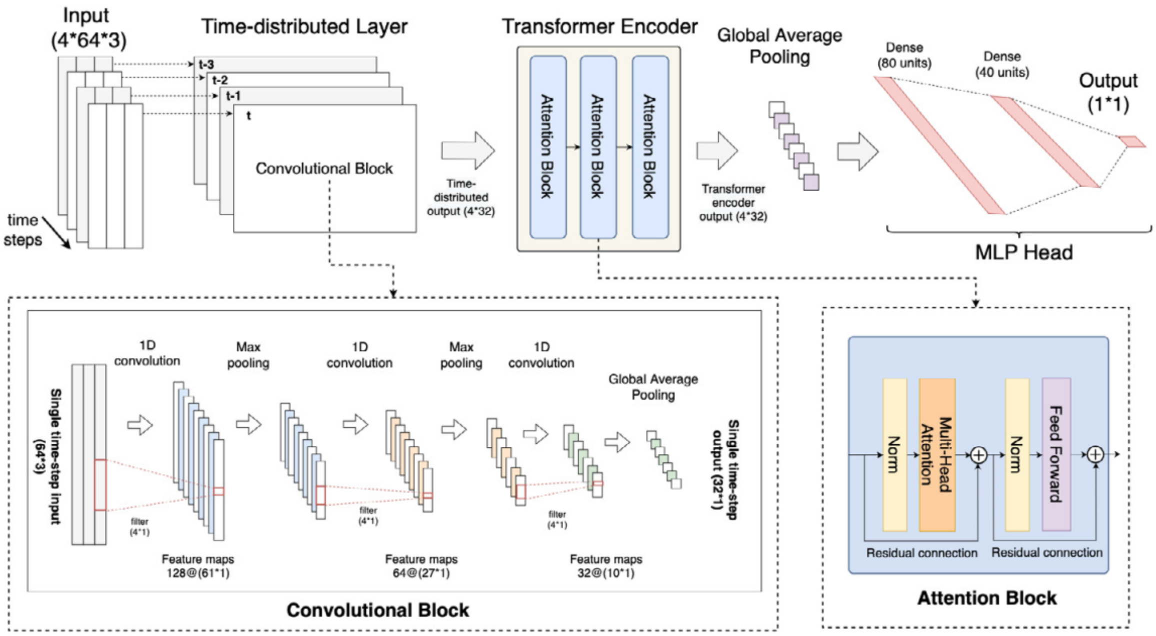


Figure 4.6: The transformer architecture designed by Sigcha et al. [8].

The convolutional backbone in this context operates on data in batches, convolving the batches of the signal with specific filters (or kernels) whose weights are optimized during training. This process generates new signals by applying the convolution operation across the entire signal and filter weights. Following this, pooling techniques such as max pooling or average pooling are often applied to reduce dimensionality and emphasize the significant features of the signal. The objective is to highlight key aspects of the signal through dimensionality reduction, retaining only the dominant aspects (see Figure 4.7).

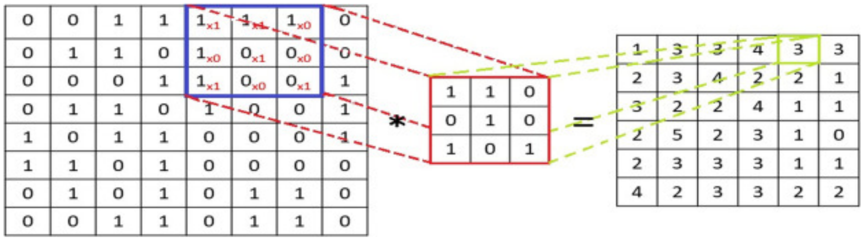


Figure 4.7: Visualization of a convolutional layer operations [9].

Moving to the neck of the model, the transformer encoder block comprises attention blocks. Each attention block takes a query token, a set of key tokens, and a set of value tokens as inputs. It calculates attention scores by performing a dot product between the query and key tokens, followed by a softmax operation to obtain normalized weights. These weights determine the importance of each key token for the given query token. The attention block then computes a context vector, which is a weighted sum of the value tokens. This context vector represents a comprehensive representation of the query token, capturing dependencies and long-range relationships within a sequence effectively. Transformers often employ multiple attention blocks, allowing the model to focus on relevant information and capture intricate dependencies, making it a powerful tool for temporal signal analysis tasks. Techniques like multi-head attention, residual connections, and layer normalization all help to further enhance the model's performance [54].

Finally, the regular neural network used as the head is the same as the neural network described earlier. Overall, the combination of a convolutional backbone, transformer encoder blocks for the neck, and a regular neural network for the head form the architecture developed by Sigcha et al., providing a powerful and versatile framework for various tasks [8].

4.3 Tuning Methodology

All models were tuned to ensure the optimal hyperparameters were used to produce the best performance for the datasets used in this work. To achieve the tuned models, a randomized grid search for specific hyperparameter options for each model was conducted. Each potential model was graded based on the macro F1-score for the FoG labels, after which the top-performing model was selected. The results for both the classical and deep network model tuning can be found in the subsequent section 5.2.1.

4.3.1 Classical Model Tuning

For the six classical models, the randomized grid search was carried out with the use of the scikit-learn framework which iterates through potential permutations and then scores each model according to a shuffled stratified 5-fold cross-validation split [52]. Since cross-validation is used within the default randomized grid search for scikit-learn, the MJFF and Sydney datasets were combined and passed altogether to the classical model tuning pipeline. Each model tunning process was run for a maximum of 250 iterations, with fewer iterations run if half the number of possible permutations was less than 250. The hyperparameters kept constant and considered during tuning for each model can be seen in Table 4.1.

4.3.2 Deep Network Model Tuning

The deep model tuning procedure differed from classical models in several ways. To tune the deep model, the Optuna Python framework was employed, as PyTorch does not provide native hyperparameter tuning options [56, 7]. Due to computational and time constraints, each possible model was evaluated only once, using the MJFF dataset for training and the Sydney dataset for testing. For efficiency, a maximum of 100 iterations was run for the possible model permutations, compared to the maximum of 250 for the classical models.

Model	Parameter	Values	Status
LR	Max Iterations	400	Constant
	Solver	liblinear	Constant
	Multi-Class	ovr	Constant
	Class Weight	balanced	Constant
	Penalty	l1, l2	Tuned
	Inverse of Regularization Strength	(0.1,100)	Tuned
RF	Class Weight	balanced	Constant
	Number of Estimators	(50, 200)	Tuned
	Criterion	gini, entropy, log loss	Tuned
	Max Features	sqrt, log2	Tuned
SVM	Max Iterations	400	Constant
	Class Weight	balanced	Constant
	Kernel	linear, poly, rbf, sigmoid	Tuned
	Regularization	(0.1, 10)	Tuned
KNN	Number of Neighbours	5,100	Tuned
	Weights	uniform, distance	Tuned
	Algorithm	auto, ball tree, kd tree, brute	Tuned
	Leaf Size	(20, 100)	Tuned
	Power Parameter for the Minkowski Metric	(1,5)	Tuned
ADT	Class Weight	balanced	Constant
	Number of Estimators	50,200	Tuned
	Criterion	gini, entropy, log loss	Tuned
	Learning Rate	(0.5,3)	Tuned
	Max Features	auto, sqrt, log2	Tuned
NN	Learning Rate	adaptive	Constant
	Early Stopping	TRUE	Constant
	Max Iterations	300	Constant
	Batch Size	64, 128, 256, 512, 1024	Tuned
	Activation	relu, tanh	Tuned
	Initial Learning Rate	0.001, 0.0001, 0.01, 0.1	Tuned
	Number of Iterations with no Change	(2, 10)	Tuned
	Hidden Layer Size	(10, 100)	Tuned
	Solver	sgd, adam	Tuned

Table 4.1: Classical model constant hyperparameters, and pool of hyperparameters optimized during the tuning process.

Additionally, a unique constraint was applied to the deep model tuning, ensuring the top-performing model had training and validation losses that converged. Convergence was checked through a visual analysis to verify both losses decreased over the epochs by at least 0.2, had limited noise (i.e., ≤ 0.1 change in loss between epochs), and finally, appeared to be approaching a steady state near the end of the epochs (i.e., the derivative appeared to be approaching zero). If neither converged or if the validation loss failed to converge, the next best-performing model was considered.

The hyperparameters considered for the deep model tuning are listed in Table 4.2. The complexity of the transformer encoder allowed numerous parameters to be altered, although the architecture parameters were limited to options that did not impact the overall size and structure of the data passed to the model. Consequently, the deep model tuning primarily focused on training parameters to improve the model's learning rather than altering the model's architecture.

Parameter	**Values**	**Parameter Type**
Optimizer	Adam, SGD	Hyperparameter
Learning Rate	0.01, 0.001, 0.0001, 0.00001, 0.005, 0.0005, 0.00005	Hyperparameter
First Dropout	(0.2, 0.6)	Architecture
Second Dropout	(0.6, 0.9)	Architecture
Batch Size	128, 256, 612	Hyperparameter
Momentum	(0, 0.3)	Hyperparameter
Number of Encoder Heads	1, 2, 4	Architecture

Table 4.2: Pool of parameters optimized during the deep model tuning process.

4.4 Test Cases

Various test cases were performed to examine the performance of the included models. The first test case involved training the models on the MJFF data and testing them on the holdout Sydney dataset. This scenario simulates a situation where a pre-trained model exists and a different dataset with a different sensor but a similar setup attempts to utilize the model. The second test case included training and testing on only the Sydney data using a LOSO format. This case represents a practical environment where an organization develops and trains a model on internal data and then applies it to a new participant.

In the second test case, the deep network model was excluded due to the lack of trials in the Sydney dataset compared to the MJFF data. This limitation would lead to high overfitting toward the training data or abnormally high bias. This test case simulates the environment of an organization utilizing an already developed model on their in-house data that has overlapping sensor sets.

The models included in these test cases were all of the tuned classical models, the tuned transformer model, and finally, the stock transformer with and without early stopping. Also, both of these test cases were executed for the current standard binary domain and the suggested ternary FoG classification (i.e., where the FoG labels are divided into akinetic and kinetic subtypes). The inclusion of the binary scenario aimed primarily at benchmarking the implementation of the tuned models and datasets by comparing the results with current literature results. If the implemented models and datasets' performance were similar to the literature, this was taken as an indication that the execution lacked any major flaws, and ternary domain exploration was granted. Little comparison, between the binary and ternary classification performance was made. Detailed results for these test cases can be found in section 5.3.

Additionally, it should be noted, models were only ever tested on the Sydney dataset whether this was as a holdout set or in the LOSO fashion. No testing was conducted on the MJFF dataset due to a large participant population, no records for the FoG subgroup and finally, the lack of motor situations annotations [47].

Lastly, an exploratory test case was conducted, involving training and testing only on the Sydney dataset using the LOSO approach, but with the incorporation of the sternum and bi-lateral feet IMU sensors. The purpose of this test case was to investigate the implications of solely using a lumbar sensor in FoG classification. As with the second test case, the deep network model was not included in this scenario due to data limitations. The results for this additional test case can be found in section 5.4.1.

4.4.1 Performance Metrics

To gauge the performance of the models during each test case, a combination of standard machine learning performance metrics and clinical FoG metrics were used. The machine learning metrics used were accuracy, sensitivity, specificity, precision and F1-score of the classification. Accuracy is beneficial to highlight the overlap of correct labels throughout all the potential classes. Sensitivity and specificity on the other hand help to represent the proportion of correct class predictions within the total number of true class instances. Sensitivity in this case is for the positive instances (i.e., FoG subtypes), whereas the specificity is for the negative class (i.e., no-FoG). Precision is similar to sensitivity but instead analyzes the proportion of correct class predictions to the total number of class predictions. Lastly, the F1-score was considered with the most weight out of these metrics due to it combining both the sensitivity and precision values into a single value which provides a balanced metric that analyzes both the false and true positives. For the ternary case, the macro averages for sensitivity, precision and F1-score were found with only the scores for the FoG labels. This was done to remove the performance bump which the non-FoG class would add to the ternary case only. This is vacant in the binary case since only a single-column binary vector is necessary to represent the predicted labels.

As for the clinical metrics, the overall FoG severity was analyzed for individual participants based on the time spent frozen with the Jerusalem protocol definitions (mild: $\%FoG < 10$, moderate: $10 < \%FoG < 50$, severe: $\%FoG > 50$). Other clinical severity tools were excluded from this analysis due to a lack of availability in the datasets, and/or complexity when implemented in an automatic sense (i.e., final severity scores require some clinical interpretations) [10, 16, 17, 18].

Lastly, strip chart plots that showcase the true and predicted labels for FoG, as well as the motor situation, were made for each trial and combined into a single figure for each participant. These figures provided an outlet for visual analysis of the performance of the models, as opposed to strictly the numerical approach. This metric was included, rather than solely relying on valued metrics because it helps to indicate the relation of the timing or onset and offset of the predicted and true labels. As well, it helps to gain a visual understanding of the performance within specific motor situations, and during the transition between situations (i.e., taking the first step after the 360-degree counterclockwise turn).

Chapter 5

Results & Discussion

5.1 Feature Engineering

5.1.1 Results

After conducting the feature engineering procedure on the initial pool of features listed in Table 3.4, the optimal number of features was determined. As a reminder, each final set of features was found to be most optimal for the average F1-score across all the non-tuned classical models after conducting the empirical study explained in section 3.4. When considering only the lumbar IMU sensor location and based on mutual information, it was found that 19 features were optimal for the binary classification and 14 features for the ternary classification, out of the original 24 features. Since only the lumbar IMU sensor location was considered, there was no need for additional features to ensure symmetry. The selected features for the binary and ternary classifications can be found in Table 5.1 and Table 5.2.

In the case where all available sensors were considered, the optimal number of features for both binary and ternary classifications was 57, out of the original 144 features. These

Feature	Sensor Types	Sensor Locations	Directions
Root mean squared	Acc.	Lumbar	AP, ML, V
Standard deviation	Acc.	Lumbar	AP, ML, V
Kurtosis	Acc.	Lumbar	AP, ML
Skewness	Acc.	Lumbar	V
Number of dominant peaks	Acc.	Lumbar	V
Number of zero crossings	Acc.	Lumbar	AP, ML, V
Extended freezing index	Acc.	Lumbar	AP, ML, V
Displacement	Acc.	Lumbar	AP, ML, V

Table 5.1: The 19 selected features based on mutual information for binary freezing of gait classification (AP - anterior-posterior, ML - medial-lateral, V - vertical).

Feature	Sensor Types	Sensor Locations	Directions
Root mean squared	Acc.	Lumbar	AP, ML, V
Standard deviation	Acc.	Lumbar	AP, ML, V
Skewness	Acc.	Lumbar	V
Number of dominant peaks	Acc.	Lumbar	V
Number of zero crossings	Acc.	Lumbar	AP, ML, V
Extended freezing index	Acc.	Lumbar	AP, ML, V

Table 5.2: The 14 selected features based on mutual information for ternary freezing of gait classification (AP - anterior-posterior, ML - medial-lateral, V - vertical).

feature sets were selected using ANOVA for both situations. The detailed list of features included in this set can be found in Appendix B.

5.1.2 Discussion

From the selected features when only considering the lumbar sensor it can be seen that all possible features are represented in at least one axis for the binary classification. As for ternary classification, only kurtosis and displacement were not represented. Additionally, it should be noted, that 4 features including root mean squared, standard deviation, number of zero crossings and extended freezing index, were represented in all three axes of the lumbar accelerometer. This hints toward these four features having a high correlation to the labels, and are likely the driving features for the classical models. However, due to the limited capabilities of the feature engineering test, and its empirical nature, there are no inferences immediately identifiable from these optimal features.

5.2 Hyperparameter Tuning

5.2.1 Results

5.2.1.1 Classical Models

After conducting the tuning procedure for the classical models, the hyperparameters for each model shown in Tables 5.3 were found to yield the highest average macro F1-score in the shuffled stratified 5-fold cross-validation split in both the binary and ternary cases. The tuned parameters for the scenario where more than just the lumbar accelerometer is input to the models can be found in Appendix B.

5.2.1.2 Deep Network Model

Similarly, upon executing the tuning procedure for the deep network model, the hyperparameters presented in Table 5.4 were identified as yielding the highest average macro F1-score on the Sydney dataset.

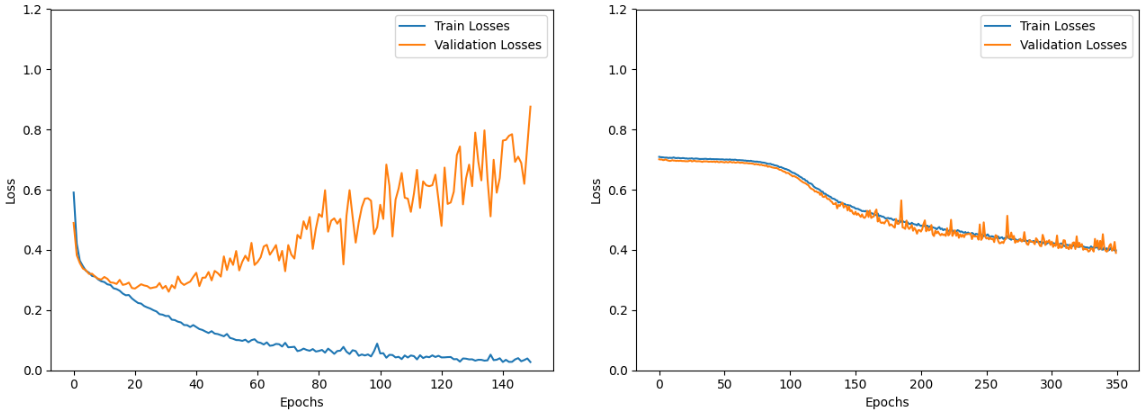
Furthermore, a comparison between the loss during training for the untuned model (utilizing Sigcha et al. stock parameters) and the tuned model is illustrated in Figure 5.1 and Figure 5.2 [8].

Model	Parameter	Binary Tuned Value	Ternary Tuned Value
LR	Penalty	l2	l1
	Inverse of Regularization Strength	6.6	3.6
RF	Number of Estimators	183	166
	Criterion	log_loss	log_loss
	Max Features	log2	sqrt
SVM	Kernel	sigmoid	sigmoid
	Regularization	8.9	8.2
KNN	Number of Neighbours	15	5
	Weights	distance	distance
	Algorithm	auto	brute
	Leaf Size	20	60
	Power Parameter for the Minkowski Metric	1	1
ADT	Number of Estimators	200	166
	Criterion	entropy	log_loss
	Learning Rate	1.026316	2.342105
	Max Features	sqrt	log2
NN	Batch Size	64	128
	Activation	tanh	tanh
	Initial Learning Rate	0.001	0.001
	Number of Iterations with no Change	8	8
	Hidden Layer Size	30	60
	Solver	adam	adam

Table 5.3: Optimal values for classical model hyperparameters during the tuning procedure for both binary and ternary classification.

Parameter	Stock Value	Binary Tuned Value	Ternary Tuned Value
Optimizer	Adam	SGD	Adam
Learning Rate	0.0006	0.0005	0.00005
First Dropout	0.25	0.2	0.4
Second Dropout	0.7	0.7	0.8
Batch Size	512	512	512
Momentum	-	0.25	-
Number of Encoder Heads	3	4	4

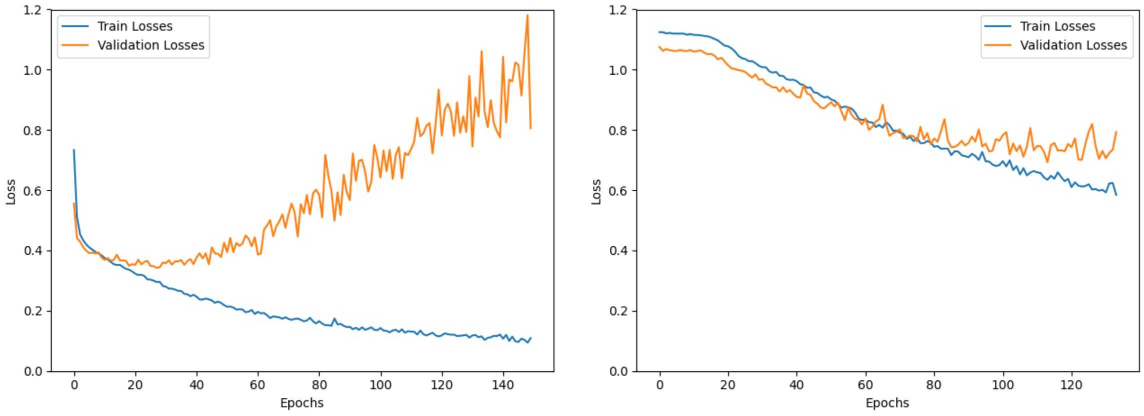
Table 5.4: Optimal values found for the deep network model hyperparameters during the tuning procedure for both binary and ternary classification.



(a) Stock Model

(b) Tuned Model

Figure 5.1: Comparison of stock and tuned deep model train and validation losses for binary classification.



(a) Stock Model

(b) Tuned Model

Figure 5.2: Comparison of stock and tuned deep model train and validation losses for ternary classification.

5.2.2 Discussion

Since no specific classical model hyperparameters were referenced as a baseline, insight gained from the tuning results is minimal. However, for the transformer model, a comparison of the stock untuned model to the tuned models can be made. From Figure 5.1 and 5.2, the model is severely overfitting to the training data with the stock parameters, indicated by the validation and training loss estimates. This is not desirable, as it will likely lead to poor performance when the model is tested on non-training data. The tuned model addressed this by modifying the hyperparameters to increase the bias and reduce the overfitting (i.e., increase the magnitude of the loss, but reduce the difference between the validation and the training loss). This is evident in Figure 5.1 and 5.2, where the validation loss function does not diverge from the training loss to the same degree as the stock model. However, as mentioned this comes at the cost of higher loss magnitude for the the training loss and the validation loss in the ternary domain. Overall, this was accomplished with slower training rates, and/or larger dropout values.

An important thing to note for Figures 5.1 and 5.2 is that there was no early stopping implemented within the stock model loss plot. This was done since Sigcha et al. did not implement any early stopping and trained for the entire 150 epochs, with no loss analysis. However, it can be observed that the final validation loss could be lower with the stock model compared to the tuned model if early stopping was implemented. As mentioned in section 4.4, along with the tuned transformer model, the stock transformer with and without early stopping were both tested for all test cases. Despite this inclusion, it was found that both the stock model with and without early stopping did not achieve high enough F1-scores to be in the top five test cases for binary or ternary classification (see Tables B.1 and B.2 in section B.1 within the Appendix for performance metrics). The poor performance of the non-early stopping variant of Sigcha's stock transformer model is likely attributed to the overfitting of the model. However, for the stock model with early stopping, the reason for the poor performance is most likely from different sources.

The binary and ternary stock transformer models both stopped around 40 to 50 epochs when early stopping was implemented, which gave validation losses of roughly 0.3 and 0.4 respectively. These are lower than the final validation losses in the tuned model while still limiting overfitting. In spite of this, both of these models performed much worse than their tuned counterparts. In particular, both stock models struggled heavily with predicting instances of FoG. This highlights a clear disconnect between the loss function, cross-entropy loss, and the main performance metric, the F1-score. This disconnect can likely be attributed to the fact that the majority of the FoG datasets are non-FoG. However, since non-FoG is the negative class, it is not included in the F1-score. Therefore, the low validation loss was created from overfitting towards the non-FoG class, instead of the positive classes. This overfitting to the negative class could happen for many reasons, with the unbalanced datasets being a large contributor, but could also be associated with the lower number of epochs and the initialization of the weights. Based on these findings it appears as though the initial weights are more optimal for the majority class, non-FoG, and thus fewer epochs restrict the extent to which the model starts to sacrifice the high specificity for more positive class instances.

5.3 Classification

Due to a large number of models and two test cases, only the results for the top five tests were considered for both the binary and ternary test cases. Since multiple test cases were run for the classical models where the training set was modified, the classical models had two opportunities to appear in the top five tests. The performance of each test was gauged primarily on the overall F1-score. Additionally, since multiple training sets were used across the top five tests, the direct comparison of these models has little validity. However, in this study, the representation of performance does not aim to compare models directly but to identify the most optimal conditions for the model architectures considered.

5.3.1 Binary Classification

5.3.1.1 Results

The performance metrics for the combined top five tests for binary classification can be found in Tables 5.5 and 5.6. Table 5.5 represents the tests within the top five performing, based on F1-score, that were trained on the Sydney dataset and tested in a LOSO fashion. Whereas Table 5.6 represents the tests within the top five performing that were trained on the MJFF data, then tested on the Sydeney as a holdout set. The test code identifier within these tables helps to distinguish the FoG classification domain, training data, and position within the top five tests (e.g., BS2 represents binary, trained on Sydney and 2nd in the top five tests, and BM1 represents binary trained on MJFF, and 1st in the top five tests).

Model	**Test Code**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-Score**
Logistic Regression	BS2	0.73	0.61	0.79	0.61	0.61
AdaBoosted Decision Tree	BS4	0.72	0.52	0.82	0.61	0.56

Table 5.5: Performance metrics for the tests trained on the Sydney dataset and within the top five F1-scores for binary tests.

Model	**Test Code**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-Score**
Transformer	BM1	0.74	0.83	0.69	0.58	0.69
Logistic Regression	BM3	0.70	0.61	0.74	0.56	0.58
AdaBoosted Decision Tree	BM5	0.69	0.47	0.81	0.57	0.51

Table 5.6: Performance metrics for the tests trained on the MJFF dataset and within the top five F1-scores for binary tests.

5.3.1.2 Discussion

From the binary classification performance, models tested on the Sydney dataset demonstrated lower-end performance, but close to the range in current literature [26]. A potential reason why many of the classical models do not fit in the sensitivity and specificity ranges stated before is the limitation of using only lumbar features. Often these more simplistic models are paired with a wider pool of descriptive features, which could hinder the performance in this test case.

Regarding the tuned transformer model, the performance on the Sydney dataset is close to the performance Sigcha et al. achieved [8]. The first potential reason for the drop in performance, mainly with specificity, could be associated with the amount of freezing exhibited in the test set. The Sydney dataset has over 30% of the data spent in a freezing episode, whereas the REMPARK dataset used by Sigcha et al. is composed of 10.5% freezing [10, 8]. This increase in the proportion of negative class (i.e., no FoG) could aid specificity by having the model slightly favour that class, which would essentially increase the specificity (i.e., recall of the negative class). The second, larger contributor, is no calibration process was carried out on the probability outputs from the model in the current study; instead, the maximum value was assumed to be the desired class. Calibration helps to further improve the performance and was utilized by Sigcha et al. when achieving their maximum F1-score of 0.71 [8]. Other contributing factors to the slight decline in performance compared to literature include the window sizes considered, FoG annotation definitions, and the heterogeneity seen with the Sydney dataset and between the MJFF and Sydney dataset.

Despite the slight drop in performance, the scores achieved by the binary models help to indicate the test setup with regard to raw data, features, labels, and architectures is functional. Therefore, the ternary FoG domain is investigated in more detail.

5.3.2 Ternary Classification

5.3.2.1 Results

The performance metrics for the combined top five tests for ternary classification can be seen in Tables 5.7 and 5.8. Similar to binary, Table 5.7 represents the tests within the top five that were trained on the Sydney dataset, and Table 5.8 represents the tests within the top five that were trained on the MJFF dataset. The test codes in these tables are formatted in a similar fashion (e.g., TS3 represents ternary, trained on Sydney and 3rd in the top five tests). Additionally, overall confusion matrices for all top five tests can be seen in Figure 5.3.

When just considering the top two test cases (TS1 and TM2), the performance broken

Model	Test Code	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Logistic Regression	TS1	0.70	0.49	0.81	0.49	0.49
AdaBoosted Decision Tree	TS3	0.66	0.37	0.82	0.47	0.41
K-Nearest Neighbour	TS5	0.62	0.22	0.83	0.29	0.25

Table 5.7: Performance metrics for the tests trained on the Sydney dataset and within the top five F1-scores for ternary tests.

Model	Test Code	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Transformer	TM2	0.59	0.47	0.66	0.42	0.42
Support Vector Machine	TM4	0.26	0.60	0.08	0.17	0.27

Table 5.8: Performance metrics for the tests trained on the MJFF dataset and within the top five F1-scores for ternary tests.

down by motor situation and freezing of gait subgroup can be seen in Table 5.9 and 5.10 respectively.

Model (Test Case)	Motor Situation	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Percent	Predicted Akin. Percent	True Kin. Percent	Predicted Kin. Percent
Logistic Regression (TS1)	Forward Walking	0.74	0.38	0.82	0.15	0.22	7.58%	18.69%	11.11%	5.05%
	Right Turn	0.63	0.45	0.83	0.65	0.53	39.67%	28.12%	13.72%	6.52%
	Left Turn	0.64	0.48	0.77	0.52	0.50	32.43%	30.92%	12.37%	8.14%
	Doorway	0.83	0.69	0.85	0.46	0.55	12.24%	18.51%	1.64%	4.63%
Transformer (TM2)	Forward Walking	0.71	0.62	0.73	0.36	0.41	7.58%	25.76%	11.11%	10.10%
	Right Turn	0.57	0.52	0.62	0.59	0.53	39.67%	29.08%	13.72%	29.62%
	Left Turn	0.48	0.45	0.51	0.41	0.40	32.43%	27.30%	12.37%	34.99%
	Doorway	0.65	0.25	0.72	0.21	0.22	12.24%	12.54%	1.64%	22.69%

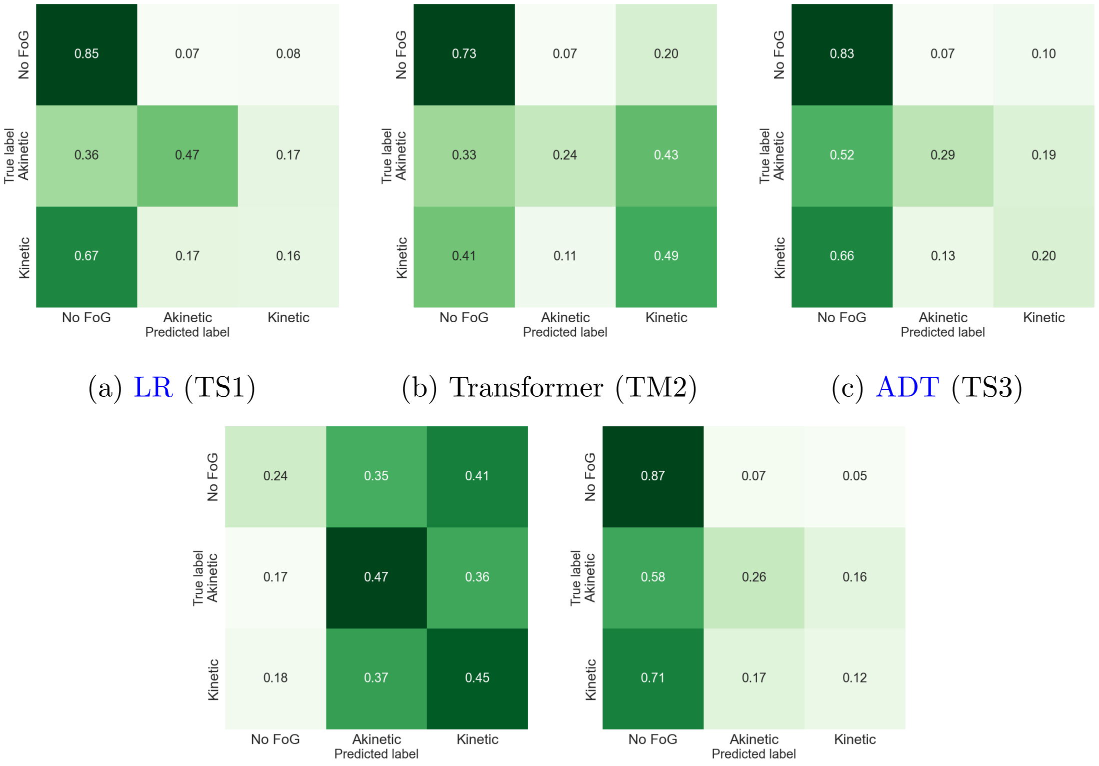
Table 5.9: Performance metrics for the top two test cases in ternary classification broken down by motor situation throughout the trials.

Model (Test Case)	FoG Subgroup	Group Size	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Percent	Predicted Akin. Percent	True Kin. Percent	Predicted Kin. Percent
Logistic Regression (TS1)	Anxiety	3	0.71	0.16	0.81	0.23	0.18	1.26%	14.75%	14.57%	9.35%
	Motor	2	0.57	0.30	0.88	0.53	0.38	39.77%	22.28%	13.60%	1.30%
	Sensory Attention	5	0.79	0.82	0.77	0.67	0.74	27.73%	33.52%	2.85%	8.26%
Transformer (TM2)	Anxiety	3	0.59	0.56	0.60	0.25	0.35	1.26%	13.13%	14.57%	32.01%
	Motor	2	0.45	0.32	0.59	0.41	0.33	39.77%	23.32%	13.60%	32.77%
	Sensory Attention	5	0.70	0.63	0.73	0.60	0.60	27.73%	27.64%	2.85%	18.71%

Table 5.10: Performance metrics for the top two test cases in ternary classification broken down by subgroup distribution across participants.

Furthermore, the breakdown by the participant for the top two models can be seen in Table 5.11 and 5.12 respectively. Similar breakdowns for the other top three models, along with the confusion matrices for individual participants can be found in Appendix B.

Lastly, after further breakdown by the participant, the strip charts aiding in the visualization of the labels of the top two test cases for each trial for participants 28FV, 76CA,



(a) LR (TS1)

(b) Transformer (TM2)

(c) ADT (TS3)

(d) SVM (TM4)

(e) KNN (TS5)

Figure 5.3: Ternary confusion matrics for the top five performing tests.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.35	0.00	0.35	0.00	0.00	Mild (0.0%)	Moderate (30.77%)	Mild (0.0%)	Moderate (34.62%)
85TL	0.96	0.00	0.97	0.00	0.00	Mild (0.0%)	Mild (3.05%)	Mild (0.76%)	Mild (0.0%)
76CA	0.77	0.24	0.81	0.07	0.11	Mild (2.49%)	Mild (6.76%)	Mild (3.56%)	Moderate (14.59%)
21DH	0.88	0.07	0.98	0.01	0.02	Mild (0.71%)	Mild (3.57%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.83	0.12	0.92	0.28	0.14	Mild (2.29%)	Mild (7.8%)	Mild (9.17%)	Mild (2.75%)
83OS	0.71	0.31	0.80	0.23	0.27	Mild (1.01%)	Mild (0.0%)	Moderate (16.58%)	Moderate (22.61%)
45PG	0.51	0.00	0.64	0.00	0.00	Mild (0.0%)	Moderate (46.76%)	Moderate (20.14%)	Mild (0.72%)
28FV	0.56	0.26	0.86	0.41	0.32	Moderate (32.5%)	Moderate (20.77%)	Moderate (16.92%)	Mild (1.17%)
97MU	0.58	0.39	0.97	0.93	0.55	Severe (64.57%)	Moderate (27.43%)	Mild (2.29%)	Mild (1.71%)
39KR	0.85	0.89	0.71	0.89	0.89	Severe (76.55%)	Severe (76.82%)	Mild (1.35%)	Mild (0.27%)
Average	0.70	0.23	0.80	0.28	0.23	-	-	-	-
SD	0.18	0.26	0.19	0.34	0.28	-	-	-	-

Table 5.11: Performance metrics for the logistic regression model (TS1) in ternary classification broken down by individual within the Sydney dataset.

97MU can be seen in Figure 5.4, 5.5 and 5.6 respectively. Again, the additional strip charts for the remaining participants can be found in Appendix B.

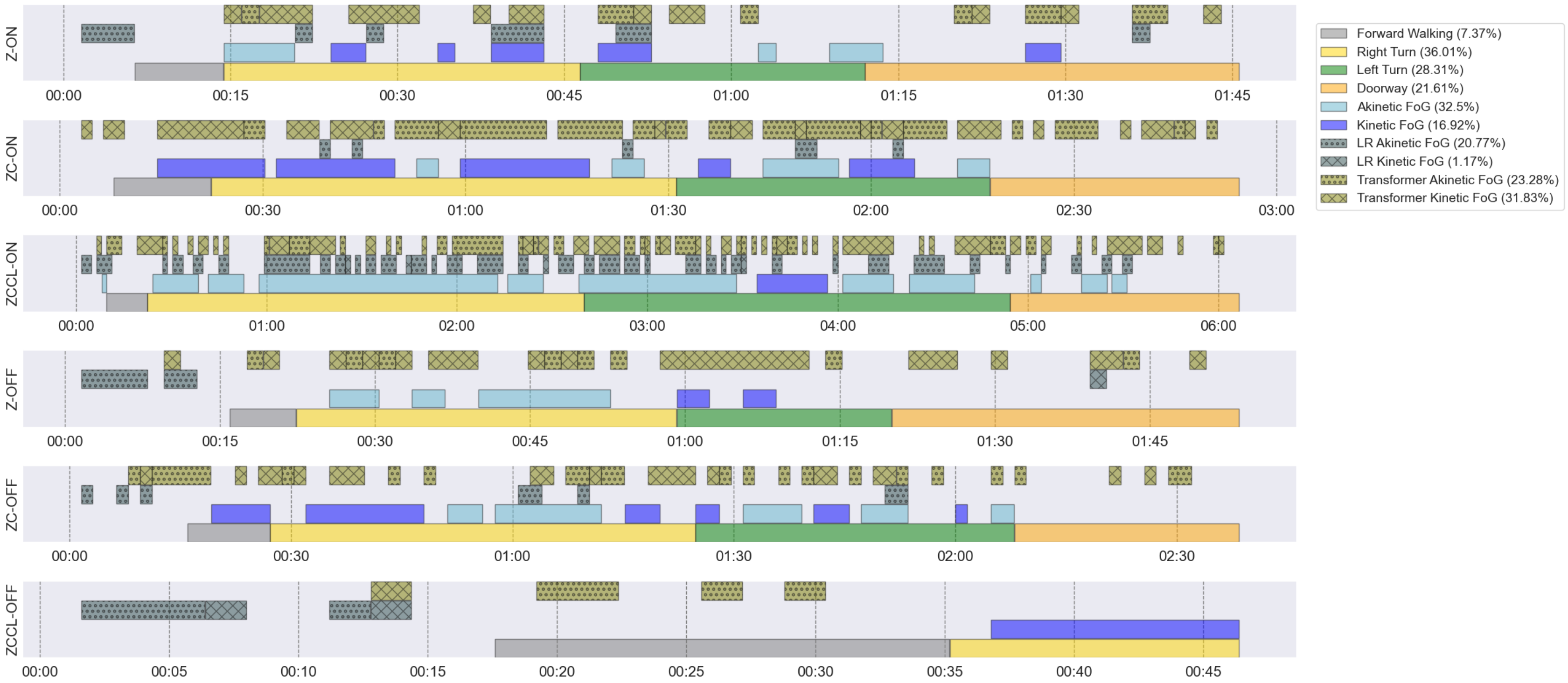


Figure 5.4: Strip chart of all true and predicted labels for each of participant 28FV's trials.

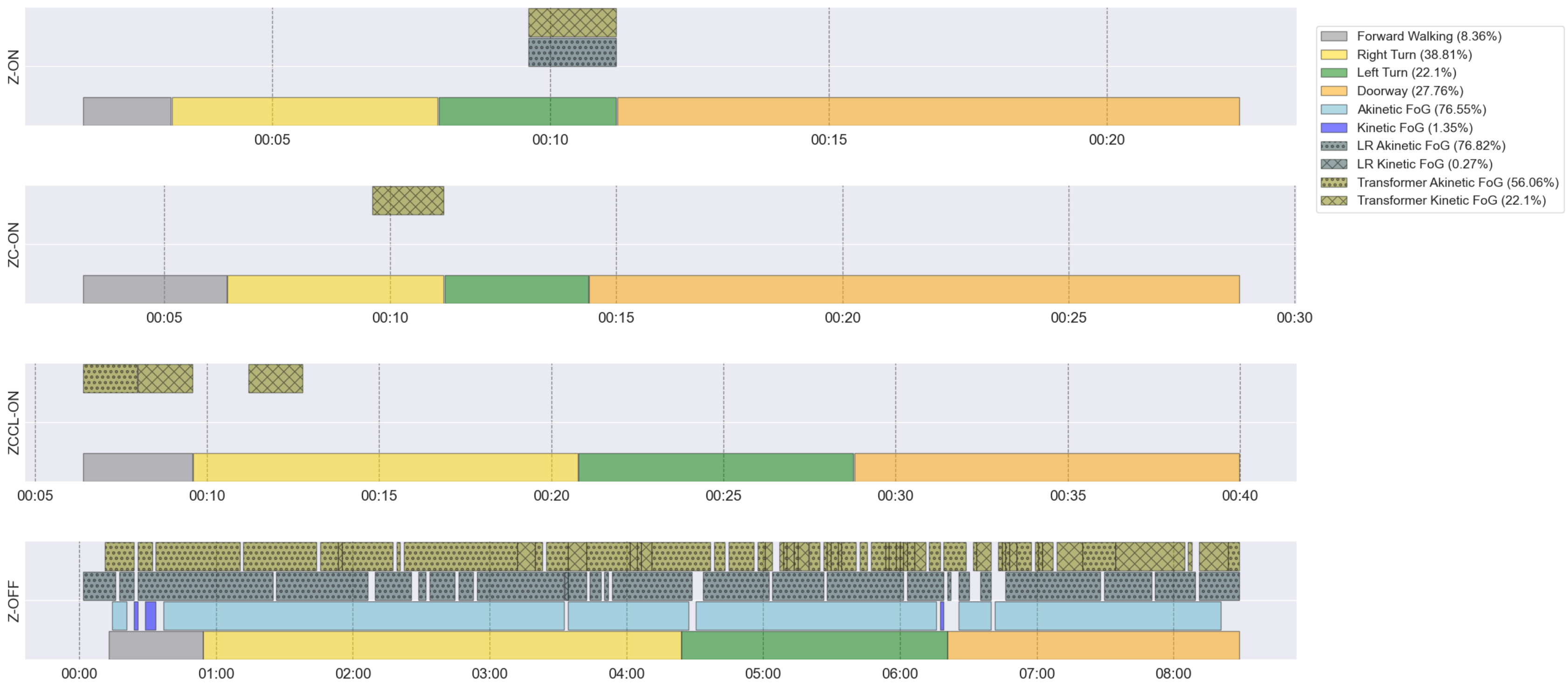


Figure 5.5: Strip chart of all true and predicted labels for each of participant 39KR's trials.

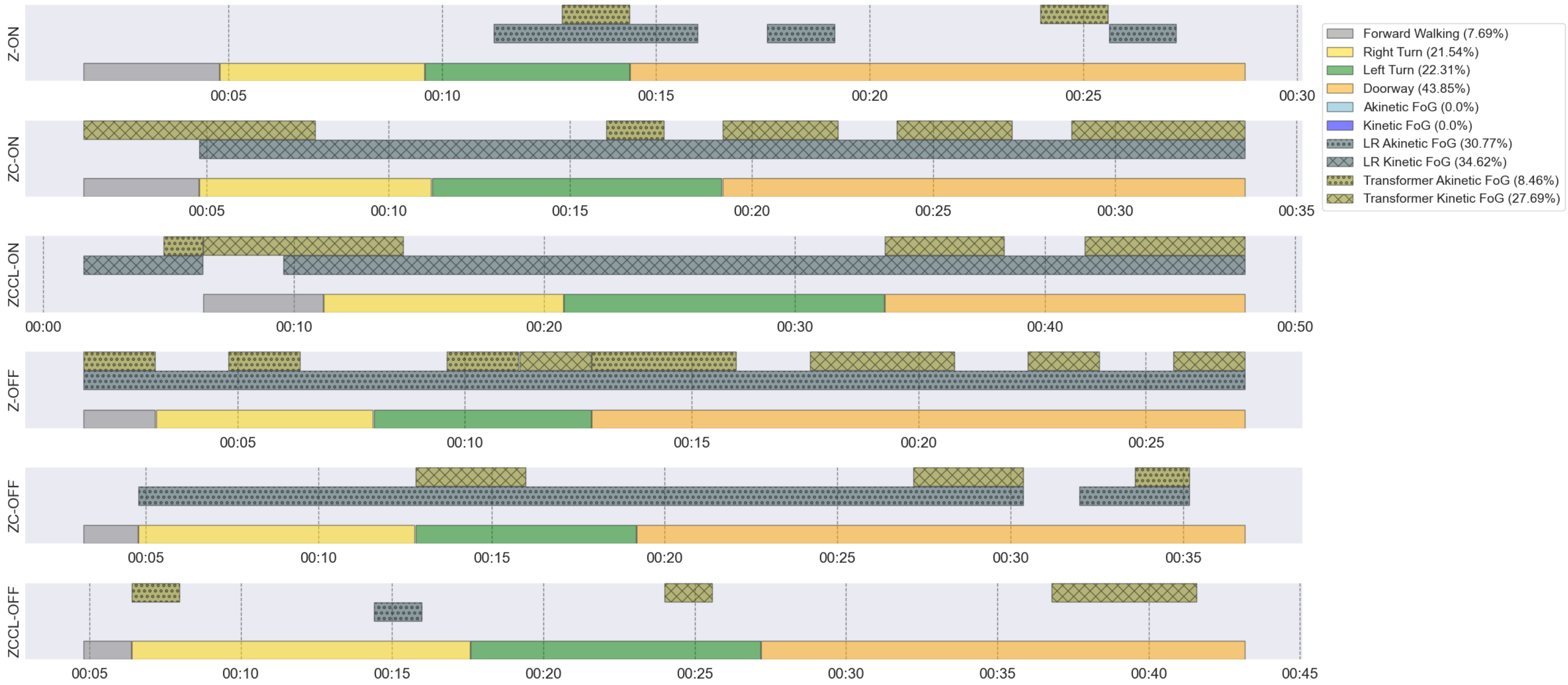


Figure 5.6: Strip chart of all true and predicted labels for each of participant 93QN's trials.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.64	0.00	0.64	0.00	0.00	Mild (0.0%)	Mild (8.46%)	Mild (0.0%)	Moderate (27.69%)
85TL	0.92	0.00	0.93	0.00	0.00	Mild (0.0%)	Mild (4.58%)	Mild (0.76%)	Mild (2.29%)
76CA	0.68	0.29	0.70	0.06	0.09	Mild (2.49%)	Moderate (17.44%)	Mild (3.56%)	Moderate (13.88%)
21DH	0.69	0.87	0.67	0.33	0.48	Mild (0.71%)	Moderate (12.14%)	Mild (10.0%)	Moderate (26.43%)
54EJ	0.63	0.36	0.67	0.15	0.21	Mild (2.29%)	Moderate (13.3%)	Mild (9.17%)	Moderate (22.48%)
83OS	0.53	0.57	0.52	0.24	0.34	Mild (1.01%)	Moderate (13.57%)	Moderate (16.58%)	Moderate (39.2%)
45PG	0.61	0.71	0.59	0.39	0.51	Mild (0.0%)	Moderate (12.23%)	Moderate (20.14%)	Moderate (36.69%)
28FV	0.42	0.31	0.54	0.31	0.29	Moderate (32.5%)	Moderate (23.28%)	Moderate (16.92%)	Moderate (31.83%)
97MU	0.53	0.35	0.88	0.90	0.48	Severe (64.57%)	Moderate (23.43%)	Mild (2.29%)	Moderate (36.0%)
39KR	0.65	0.64	0.68	0.88	0.74	Severe (76.55%)	Severe (56.06%)	Mild (1.35%)	Moderate (22.1%)
Average	0.63	0.41	0.68	0.32	0.31	-	-	-	-
SD	0.13	0.27	0.13	0.31	0.23	-	-	-	-

Table 5.12: Performance metrics for the transformer model (TM2) in ternary classification broken down by individual within the Sydney dataset.

5.3.2.2 Discussion

When observing the top five tests for the ternary classification, it can be seen that the best-performing tests only have a slight overlap with the binary results, with LR and ADT being trained on the Sydney, and the transformer trained on the MJFF still being present. However, the LR and ADT trained on the MJFF data lost enough performance to be removed from the top five tests, and be replaced with SVM trained on MJFF, and KNN trained on Sydney. As well, the logistic regression model leads in performance, whereas the transformer model drops to the second-best model.

Regardless of the test cases appearing in the top five, the performance of each dropped by roughly 20 to 30%. Based on the confusion matrices in Figure 5.3, two trends were observed with the deep and classical models, respectively. With the transformer model, it appears the model is often predicting a kinetic label when a kinetic FoG is not occurring. This is shown with 43% and 20% of the akinetic FoG events and no FoG events labelled as kinetic, respectively. Despite this large number of false positives for the kinetic class, the model still struggles with detecting kinetic properly, with only correctly classifying 49% of the kinetic freezes.

As for the classical models, they appear to be on the other end of the spectrum with the model substantially reducing the number of positive instances of the kinetic class. An interesting consequence of the reduction in kinetic class predictions is the increase in akinetic class performance. This hints towards a conflict between the kinetic and akinetic classes during prediction. This potential conflict could be a byproduct of the FoG definitions (e.g., an akinetic bout of FoG being labelled as akinetic even if there is a small amount of movement that the sensors might detect), highlighting the need for a concrete definition and/or refinement in the current definition for onsets and offsets of each FoG subtype.

One exception to these two observed patterns is the SVM model, which drastically overpredicts both akinetic and kinetic. Based on the scores and confusion matrices (see Tables 5.7 and 5.8, and Figure 5.3), the SVM model prioritized the sensitivity to result in a high F1-score which created a model which labels the majority of the data as a freezing event. Also, despite the overestimation of freezing events by the SVM model, the model does not favour either the akinetic or kinetic freezing drastically more than the other compared to the 4 other models. This indicates the SVM model is capable of extracting quantifiable differences between the two types of freezing but at the cost of poorer specificity and precision.

Lastly, the observed differences between the transformer model and the classical models with the prior favouring kinetic freezes and the later favouring akinetic freezes could have been the result of the training procedure and features used. Since the transformer model utilized signals in the frequency domain, the better performance for kinetic FoG is unsurprising, since kinetic FoG events are often paired with tremors more likely to be detected using frequency domain inputs. With the classical models, only the freezing index feature (ratio of the freezing powerband to the gait power band) directly captures information from the frequency domain, which may have resulted in the lack of performance with kinetic freezes. Specifically, the freezing index can appear to be large, which indicates a dominant freezing band, during periods of little movement due to sensor noise [27]. This suggests that the feature set selected for the tri-axial lumbar accelerometer for ternary classification is more robust for akinetic freezing rather than kinetic. Furthermore, it would be beneficial to investigate the performance of the transformer and classical models with a more balanced set of frequency and time domain data (i.e., more frequency domain features for the classical models less impacted by noise, and use of time domain signals along with frequency domain for the transformer).

Classical Model

To gain a better understanding of the performance of the classical models, the top classical model test case, TS1 with LR, was analyzed by breaking up the test set into various groups. As shown in section 5.3.2.1, the breakdown included the motor situation, FoG subgroup, and participant.

Based on these breakdown results, a similar observation can be made to the overall analysis where the LR model performs better when akinetic freezing is the dominant FoG subtype. This trend is evident in the motor breakdown (see Table 5.9), where the model performs best for right turn, left turn, and doorway motor situations. These motor situations are all dominantly akinetic freezing. In contrast, forward walking is dominantly kinetic freezing and sees a performance drop compared to akinetic freezing. This correlation with a drop in performance and lack of akinetic freezing gives hints towards the LR not having particular trouble with certain motor situations, but rather the types of freezing in those motor situations. However, there is also the possibility that the drop in performance is the byproduct of certain attributes, such as only linear movement, within the forward walking motor situation.

Additionally, a similar pattern can be observed in both the subgroup breakdown (Table 5.10) and the participant breakdown (Table 5.11), with the LR model performing best when akinetic freezing is the dominant type and when there is little kinetic freezing. For example, participant 45PG in Table 5.11 has only kinetic freezing and the model performs very poorly by classifying almost half of their trials as akinetic. It should also be noted this pattern is not always consistent; for participant 28FV, akinesia is the dominant FoG type, but the scores are only slightly higher than for participant 83OS. This lower score from 28FV could be the result of the kinetic portion bringing down the scores, as there is a moderate amount of this class. Regardless, this observation highlights a key limitation apparent with the LR model, it has great difficulties capturing kinetic events with the current information being passed into the model.

Another observation that can be made for the LR model is the reduced performance on milder freezers. This is difficult to represent for the motor situation and subgroup breakdowns but is evident in the participant breakdown (see Table 5.11). For certain participants, such as 76CA, 85TL, and 93QN, the LR model struggles greatly with being able to classify very few true positive instances of FoG, as shown through their F1-scores. Participants 76CA and 93QN have the worst performance for these milder freezers, with lower accuracy and specificity scores, indicating the model's inability to predict FoG events with proper temporal accuracy. Participant 85TL performed better, with only a few false positives compared to the other individuals, resulting in higher accuracy and specificity scores.

Overall, it can be said while the LR model is the best performing classical model, it is only acceptable for bout classification for participant 39KR. This observed performance is likely a byproduct of the features selected for the classical models. As mentioned, the feature set is from parameters in the time domain, with only the extended freezing index being in the frequency domain. This lack of frequency information could be a contributing factor to the model's limited ability to predict kinetic freezes. Based on the performance of participants such as 45PG, it seems the model is still detecting freezing events, but is fitting them to the akinetic class rather than the kinetic class. The poor performance on milder freezers could be attributed to the model overfitting to stopping events, such as voluntary stops, when there is a lack of freezing or akinetic freezing in particular.

On the other hand, performance when estimating severity is very good across participants. More specifically, the LR model predicts the correct severity for seven out of ten individuals for akinetic freezing and for six out of ten individuals for kinetic freezing. Across all participants, the akinetic severity classification achieved an F1-score of roughly 76%, whereas the kinetic severity classification achieved an F1-score of 60%. An additional benefit of the severity estimates is error margins are within a neighbouring group (i.e., no severe freezers were labelled mild, and vice versa). This better performance in severity estimation of each class for the LR model indicates that the model is labelling freezing, but the exact timing of the events rarely aligns with the true labels. This could potentially be because the classical models are only capable of considering a single window at a time, making it more difficult to understand the surrounding conditions for correctly labelling the onset and offset of a freezing bout. Also, this trouble with the exact onsets and offsets of the freezing bouts could be another result of having no concrete clinical definition for the onset and offset of a freeze, and hint toward the need for refinement of the current definitions [13].

Lastly, the performance of the severity estimations should be taken separately from the automatic labelling performance, with no comparison between them. The severity estimates are much more granular compared to the individual event labels and are a potential contributor to the increase in performance.

Deep Network Model

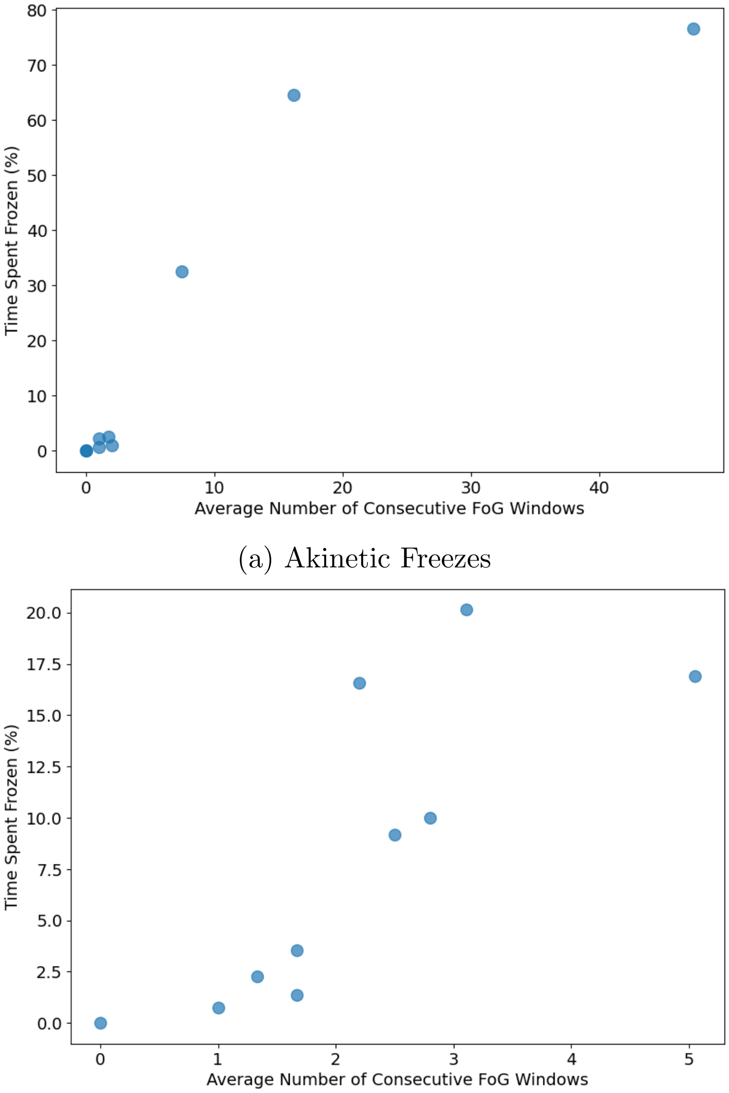
Similar to the classical models, the deep model test case within the top five, TM2 with transformer, was analyzed by the motor situation, FoG subgroup, and participant. As opposed to the top classical model, the transformer appears to be less biased towards a specific freezing type when breaking the performance down by specific groups. Specifically, the overall confusion matrix for the transformer model showed kinetic freezing being the favourite class out of the two FoG types. However, this trend is not followed when breaking the analysis down by the motor situation, FoG subgroup, and participant.

Particularly, for motor situations (see Table 5.9), the transformer model does not perform best for a single situation that is dominantly kinetic freezing. Instead, it performs best for right turn which has over double the amount of akinetic freezing compared to kinetic freezing. During the turning, left and right, the transformer does perform worse for specificity indicating it has trouble identifying periods of non-FoG during the turning movements. This could be potentially attributed to the increased number of voluntary pauses during the turning as opposed to the linear motor situations. However, due to a large amount of freezing, this does not reduce the F1-scores to below the values seen for forward walking and doorway. Also, from the subgroup breakdown (see Table 5.10), it was observed the transformer model does not perform the best for the group that is dominantly kinetic freezing, further highlighting the original hypothesis of the transformer model favouring the kinetic FoG class does not apply when subdividing the participants.

These findings indicate the lack of kinetic freezing alone does not reduce the performance of the transformer model, and the input data is capable of describing both FoG subtypes, as opposed to the difficulties seen with the classical models and the feature set. A possible explanation may lie in specific behaviours within freezes exhibited during certain motor situations or by certain FoG subgroups that confuse the model into classifying it as kinetic freezing. For example, sensory attention freezers may manifest akinetic events that more closely resemble a full akinetic freeze (i.e., no movement at all), whereas the other subgroups may include slight movement or tremors during their akinetic events (but are still clinically defined as akinesia).

When taking a look at the performance of the transformer model broken down by participant, the previously observed trend from the overall performance also appears to falter. Akin to the LR model, the only participant with acceptable scores was 39KR, and this participant actually has the most amount of akinetic freezing in the entire population. This might hint at the transformer performance not necessarily being tied to the dominance of kinetic freezing but rather the total amount of freezing. This dependency is likely the result of the deep network model considering 4 windows at a time in a single sequence, and individuals that exhibit more severe freezing for both akinetic and kinetic often have longer freezing bout durations for each subtype. This relationship can be observed by the positive correlations between the number of consecutive FoG windows and time spent frozen for both akinetic and kinetic freezing (see Figure 5.7). These longer bouts will help the sensor data to reach a steady state throughout the entire sequence passed to the model, which in turn could help diminish the number of false positives generated from labelling akinetic freezing as kinetic or vice versa.

It should be noted some participants do in fact follow the trend observed from the overall performance with the next two best-performing participants (45PG and 21DH) having mainly kinetic freezing events. However, for both of these participants, there is still a large amount of akinetic freezing predicted, further solidifying the fact the transformer model



(a) Akinetic Freezes

(b) Kinetic Freezes

Figure 5.7: Relationship between the number of consecutive FoG windows and the time spent frozen for akinetic and kinetic freezing across the ten participants in the Sydney dataset.

has difficulty distinguishing between the subtypes of freezing.

As for the severity estimates from the transformer model, the estimates for each individual freezing class are worse than the LR model. However, when combining the labels for total severity (i.e., akinetic and kinetic percentages combined), the model was able to correctly classify seven out of ten individuals, with an F1-score of 58%. However, this is not specific to the ternary domain, and would likely perform better in the binary classification since the akinetic and kinetic labels would be competing, as seen in the ternary case. As well, it should be noted again that this severity classification performance should not be directly compared to the automatic labels due to differences in the granularity of the predictions.

These findings regarding the overall severity estimate suggest the transformer model suffers more from the inability to distinguish the types of freezing. This difficulty in distinguishing between the types of freezing could be a byproduct of the use of multiple window sequences for a single sample inference. The model might be overcorrecting itself, creating a rapid switch between the labels when the data is not able to achieve a steady state. Additionally, this limitation could come down to a lack of post-calibration done for the probability outputs, as well as the inclusion of only the frequency domain as input data.

Visualization

Taking the analysis a step further, Figures 5.4 to 5.6 help illustrate the performance of the top two models broken down by trial and motor situation for a subset of participants. These plots present the predicted and true FoG labels, along with the motor context labels in a strip chart fashion. Each subfigure represents a different trial completed by the participant with the following codes for the y-axis labels: 1) “Z” is regular Ziegler, 2) “ZC” represents the Ziegler task while carrying a tray, 3) “ZCC” represents the Ziegler task while carrying a tray and completing a computational task, 4) “ON” represents the on medication state, and finally, 5) “OFF” represents the off medication state. Additionally, the legend on the right side of the Figures illustrates the hatches and colours of the bars, as well as their percentage overall for that specific individual’s trials (i.e., the percentage represents the total time the label was annotated across all trials).

Examining the first trial breakdown, Figure 5.4, for participant 28FV, a moderate freezer overall, and moderate for both akinetic and kinetic freezing. Right away, it can be seen the transformer model predicted quite a few more FoG bouts compared to the LR model, resulting in many false positives. On the other hand, the LR model is conservative, missing quite a few true FoG events, but when it does predict freezing, it often aligns well with the true labels. Almost all of the events the LR model predicts are akinetic, following the trend observed from the earlier breakdowns. As for the transformer model, it appears to rapidly switch between the akinetic and kinetic classes when labelling bouts, which aligns with the earlier hypothesis of difficulty distinguishing the FoG types.

Moving to the second trial breakdown, shown in Figure 5.5, for participant 39KR, an individual who did not complete all six trials due to large amounts of freezing in their "OFF" state. From this figure, while in their "ON" state they experienced no freezing, but during their first instance of the "OFF" state they experienced copious amounts of freezing. In particular, this individual experienced elongated bouts of akinetic freezes, likely contributing to the excellent performance of both models. With longer bouts, the data can hit a steady state, which helps the logistic regression model since it only considers one window at a time. The transformer model also performs better with regard to akinetic classification during these longer bouts. The issue of rapid changes from akinetic to kinetic predicted labels still persists but to a lesser degree. The switching in this instance could likely be attributed to movement within the akinetic freeze, but not enough to have it labelled as a kinetic freeze during the annotations by Goh et al. [10]. Yet another instance where the definition of the onset and offset of each FoG subtype could potentially be refined. Lastly, for this individual, both models perform very well for the trials with no freezing, with only a few instances of false positives. This indicates this individual has very unique freezing characteristics, easily distinguishable from regular gait patterns and voluntary stopping.

Finally, the last participant included in the trial breakdown was 93QN, another instance of the deep heterogeneity observed in the Sydney dataset population, see Figure 5.6. This individual exhibited no freezing events throughout all six trials. However, both models predicted many freezing bouts, resulting in the worst performance from both models across all participants. In particular, the LR model seems to predict long periods of akinetic and kinetic freezing, despite continuous movement by the individual. The transformer model performs slightly better in this case but still predicts multiple instances of freezing for every trial. This poor performance indicates severe differences in the gait patterns of this individual compared to the others included in the training set. These differences could include festination patterns or even slight hesitation in movement, not resulting from freezing but from other cognitive tasks, such as the counting task. Additionally, for this individual, the doorway situation made up the majority of their trials. This, in combination with numerous false positives during the doorway situation, indicates the models struggle with a reduction in pace or potentially variable pace when compared to the training population. In this situation, per-participant calibration could be beneficial, where a few benchmarking trials could be collected to calibrate the model outputs and better fit the model on a per-participant basis.

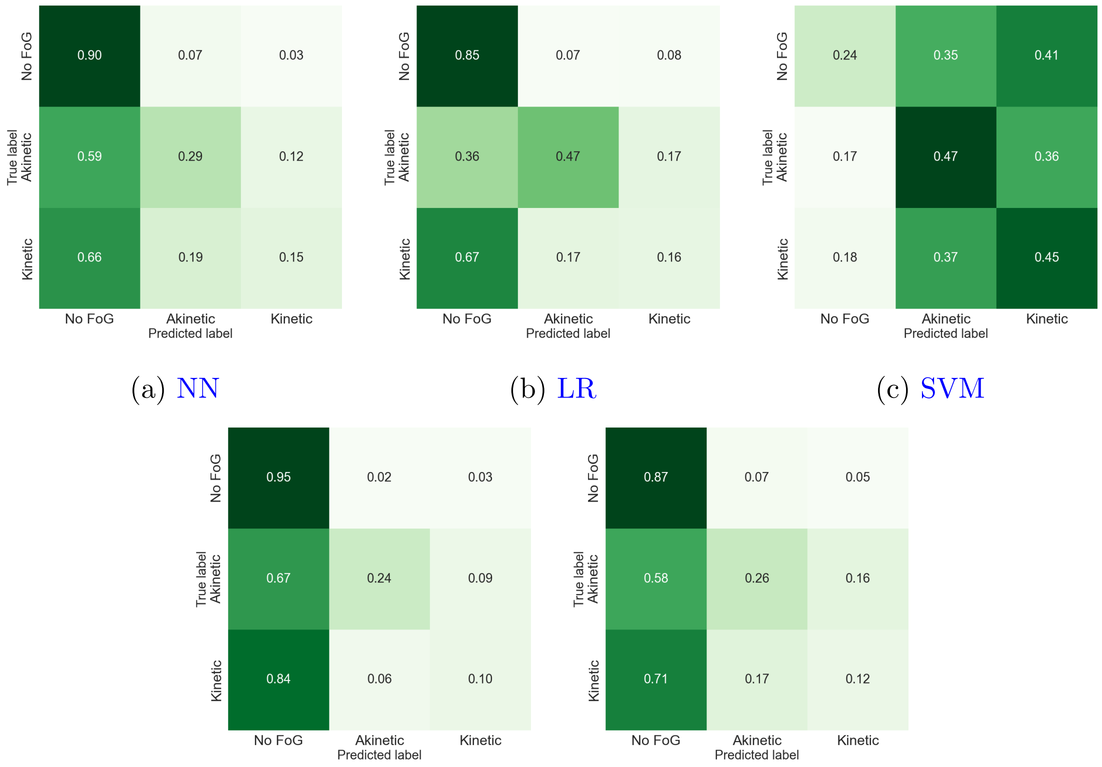
5.4 Sensor Expansion

5.4.1 Results

When considering the other sensors along with the lumbar sensor the performance for the top five models shown in Table 5.13 and Figure 5.8 were found. No test codes were used for the sensor expansion due to only the Sydney dataset being utilized.

Model	Training Data	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Neural Network	Sydney	0.77	0.55	0.89	0.58	0.54
Logistic Regression	Sydney	0.73	0.52	0.84	0.55	0.53
Support Vector Machine	Sydney	0.59	0.56	0.60	0.60	0.53
Random Forest	Sydney	0.76	0.38	0.96	0.57	0.45
K-Nearest Neighbour	Sydney	0.71	0.36	0.90	0.49	0.42

Table 5.13: Performance metrics for the top five models for ternary freezing of gait classification when considering all sensors.



(a) NN

(b) LR

(c) SVM

(d) RF

(e) KNN

Figure 5.8: Ternary confusion matrics for the top five performing models when considering all sensors.

Additionally, the performance for the second top model, LR, broken down by participant can be seen in Table 5.14. The performance of the second top model was included for comparison to the LR model when only using the lumbar accelerometer.

5.4.2 Discussion

When considering all sensors instead of just the lumbar accelerometer, the overall scores of the models exhibit a slight increase, with the top model now being the NN. Figure 5.8 shows the error patterns observed during the lumbar ternary classification persist when

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.96	0.00	0.96	0.00	0.00	Mild (0.0%)	Mild (3.85%)	Mild (0.0%)	Mild (0.0%)
85TL	0.97	0.00	0.98	0.00	0.00	Mild (0.0%)	Mild (2.29%)	Mild (0.76%)	Mild (0.0%)
76CA	0.86	0.41	0.89	0.07	0.13	Mild (2.49%)	Moderate (13.88%)	Mild (3.56%)	Mild (0.0%)
21DH	0.89	0.00	1.00	0.00	0.00	Mild (0.71%)	Mild (0.0%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.89	0.00	1.00	0.00	0.00	Mild (2.29%)	Mild (0.0%)	Mild (9.17%)	Mild (0.0%)
83OS	0.79	0.00	0.96	0.00	0.00	Mild (1.01%)	Mild (4.52%)	Moderate (16.58%)	Mild (0.0%)
45PG	0.80	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Moderate (20.14%)	Mild (0.0%)
28FV	0.61	0.58	0.63	0.45	0.49	Moderate (32.5%)	Moderate (49.41%)	Moderate (16.92%)	Mild (9.05%)
97MU	0.34	0.01	1.00	0.97	0.02	Severe (64.57%)	Mild (0.57%)	Mild (2.29%)	Mild (0.0%)
39KR	0.90	0.94	0.76	0.91	0.92	Severe (76.55%)	Severe (79.25%)	Mild (1.35%)	Mild (0.0%)
Average	0.80	0.19	0.92	0.24	0.16	-	-	-	-
SD	0.18	0.32	0.12	0.37	0.29	-	-	-	-

Table 5.14: Performance metrics for the logistic regression model in ternary classification when considering all sensors, broken down by individual within the Sydney dataset.

considering all sensors. While this performance increase while maintaining the same error patterns is promising, it also suggests expanding the feature pool could potentially extract more information and better overall performance. However, breaking the performance of the LR model down by participant reveals a more nuanced picture. The boost in performance comes from a select few participants, while others experience a drop in performance. For instance, participant 39KR consistently achieves the best performance and attains an F1-score above 90%, while participant 97MU's F1-score drastically declines from the 50% range to a score of zero, indicating no true positives were classified for any of the trials (see Table 5.14).

The discrepancy in performance is particularly intriguing for participant 97MU who exhibits large amounts of akinetic freezing. Throughout all of 97MU's trials, the LR model does not label any windows as kinetic despite the error patterns still showing better performance for akinesia across all participants. Notably, the sensor set demonstrates an improvement in severity estimates at the per-participant level, successfully classifying akinetic and kinetic severities in eight out of ten cases for both classes. This indicates the expanded sensor set aids in differentiating between akinetic and kinetic movements overall but still struggles with temporal accuracy when classifying individual freezing bouts. Lastly, this performance boost for severity estimates comes at the cost of extremely poor performance for a subset of participants, such as 97MU, whose FoG type severities were not properly classified. Therefore, while incorporating additional sensors shows promise in certain aspects, it also underscores the need to address the variability in model performance when analyzing a heterogeneous group of freezers.

Chapter 6

Conclusions & Recommendations

In this thesis, the objective was to investigate the ternary domain FoG classification with wearable sensors through current classical and deep machine learning architectures. This investigation had the following sub-objectives: 1) identification of suitable FoG datasets that can be used for training and testing of machine learning models, 2) selection of both classical and deep machine learning models based on binary classification in current literature, and optimizing the hyperparameters based on the selected datasets, 3) evaluation of model performance on a holdout set of participants, and the impact motor situation and subtype of the freezer have on the performance, and finally 4) the identification of potential applications of the ternary models based on the observed performance.

The Ziegler and MJFF datasets were deemed to be suitable for this application with both having ternary FoG labels and an overlapping sensor set of a singular tri-axial accelerometer. With these datasets, a set of seven different models based on current literature were tuned, trained and tested to provide performance metrics in the binary and ternary domains. All testing was completed on the Sydney dataset because of the motor context and FoG subgroup labelling throughout this dataset.

The performance for the binary domain was on the lower end but still aligned with what is currently observed in the literature, indicating that the implementation of the signal processing and models was conducted to permit qualitative and (limited) quantitative comparisons.

Moving on to the ternary domain, the results of the top five models indicate with the current architectures and feature pools, accurate automatic labels are not possible for all individuals. When testing the models on all participant's data at once, the majority of the classical models were observed to be biased toward akinetic freezing, while the deep model was biased toward kinetic freezing. When breaking the test scenarios down further, it was discovered that the top classical model, LR, struggled in dominantly kinetic scenarios, and also struggled with temporal accuracy. As for the deep model, it did not follow the trend of performing better in cases where the freezing was dominantly kinetic. Instead, the deep model struggled more with the differentiation of the two subtypes. The drop in akinetic performance is likely attributable to the challenge of differentiating types of freezing, especially considering the larger amount of akinetic freezing throughout the test set. In specific situations, such as large amounts of akinetic freezing for the classical models and large amounts of freezing for the deep network models, the top-performing models excelled. However, this was a minority of test cases and often F1-scores fell below 40%.

Moving away from the classification of individual bouts, and focusing purely on severity, the model's performance did improve. The top classical model, LR, was able to predict the majority of participants' akinetic and kinetic severity separately. The transformer model did not perform as well for the individual subtype severities but did perform well when predicting the overall severity (i.e., akinetic and kinetic severity combined).

6.1 Recommendations

6.1.1 Applications

Based on the observed scores of the top two models, it can be concluded the automatic ternary FoG classification using current feature pools and architectures is yet to be sufficiently accurate for clinical use. However, this does not render the output of these models useless; rather, it highlights that inferences produced by the ternary FoG classification models should not be solely relied upon for accurate automatic labels. One potential application of these models' outputs is the use of severity predictions through the percent time spent frozen, instead of the exact timing of each bout. The severity predictions are on a much coarser scale compared to the individual bout labelling and showed higher performance compared to the actual bout estimation with the top-performing model being able to classify the akinetic severity for seven out of ten participants, making this classification more valuable for practical applications. For instance, these severity predictions could be employed for long-term severity tracking, helping to identify whether an individual's symptoms are overall improving or worsening. This information, in turn, could aid clinicians in determining the appropriate treatment direction for each patient. While the severity estimates are not flawless, implementing per-participant calibration through benchmark examinations could potentially mitigate errors observed for some individuals. With this in mind, this free-living severity tracking should still not be independently used, but paired with regular in-clinic checks to help provide more insight into an individual's disease progression in free-living. This combination of in-clinic assessments and continuous monitoring through the models' predictions could enhance the management of FoG in patients.

Another potential application lies in utilizing the FoG prediction output as an annotation tool. The model could initially process the data, highlighting areas of interest clinicians might want to focus on more attentively during their analysis. While the models' performance might not be strong enough to work independently of clinicians, they could still streamline the annotation process by pinpointing potential FoG instances for further inspection.

6.1.2 Improvements

Building on the performance characterized during this work in the ternary FoG classification space, the investigation of possible solutions should be continued. One promising approach is to expand the quantity of data used and conduct a deeper exploration of model architectures and hyperparameters. An essential next step would be to characterize the performance of calibrating the model to various groups and relate it back to the non-calibrated models. Along with calibration, it may be beneficial to combine the models discussed into a majority voting system to attenuate unique error patterns observed in certain models.

The findings from the thesis suggest specifically examining voluntary stopping as a key mechanism challenging the automatic detection of akinetic freezes. In this future investigation, an additional metric that should be included is the overlap of akinetic freeze predictions with voluntary stopping labels, along with the true akinetic FoG labels. This analysis would provide a better understanding of the frequency with which the model might be misled by the lack of movement despite it being voluntary.

Moreover, the performance boost achieved by including the bilateral feet and sternum sensors, along with angular velocity, suggests lumbar sensors alone might not be sufficient for FoG detection, regardless of model complexity. Therefore, future FoG studies should explore other sensor modalities and locations, as well as larger feature pools, to gain a deeper understanding of the most informative features for FoG classification. Additional sensor modality options include an electrocardiogram (ECG) to capture heart rate, which has been shown to stay elevated during an involuntary stop (akinetic freeze) compared to dropping instantly for voluntary stops among a population of freezers [27]. This could aid in the classification of akinetic freezing. Furthermore, upper limb sensors could offer valuable free-living context by analyzing armswing waveforms and asymmetry. They may also help in detecting periods of lack of balance, which are often correlated with freezing events [57]. This inclusion of additional sensors could be used to pad the feature pool for the classical models, and also help to improve the freezing type differentiation within the transformer model.

Moving away from potential improvements of the classification models, based on the performance in the ternary domain, further investigation into FoG definitions is warranted. In particular, working towards defining quantitative thresholds for the onset and offset of both akinetic and kinetic freezing would be desirable. The transformer model struggled with movements within an already established akinetic bout, thus, revisiting the definitions for when an akinetic or kinetic FoG transitions to another type or ends is necessary. These thresholds could potentially be based on features such as displacement and velocity from the common lumbar accelerometers, or features from other sensor modelaties.

In conclusion, continuing the exploration of these aspects in FoG detection can lead to improved classification models and a more comprehensive understanding of this challenging symptom in various contexts.

References

[1] T. Reches, M. Dagan, T. Herman, E. Gazit, N. A. Gouskova, N. Giladi, B. Manor, and J. M. Hausdorff, "Using wearable sensors and machine learning to automatically detect freezing of gait during a fog-provoking test," *Sensors (Basel, Switzerland)*, vol. 20, pp. 1–16, 8 2020.

[2] D. Zoetewei, T. Herman, M. Brozgol, P. Ginis, P. C. Thumm, E. Ceulemans, E. Decaluwé, L. Palmerini, A. Ferrari, A. Nieuwboer, and J. M. Hausdorff, "Protocol for the defog trial: A randomized controlled trial on the effects of smartphone-based, on-demand cueing for freezing of gait in parkinson's disease," *Contemporary Clinical Trials Communications*, vol. 24, p. 100817, 12 2021.

[3] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 4 2016.

[4] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[5] M. A. Yusuf, M. K. Khan, T. Mahmood, M. Umer, and R. U. Afridi, "Analyzing the impact of forest cover at river bank on flood spread by using predictive analytics and satellite imagery," *International Journal of Advanced Computer Science and Applications*, vol. 10, pp. 232–239, 51 2019.

[6] M. Borowski and K. Zwolińska, “Prediction of cooling energy consumption using a neural network on the example of the hotel building,” *Proceedings 2020, Vol. 58, Page 21*, vol. 58, p. 21, 9 2020.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[8] L. Sigcha, L. Borzì, I. Pavón, N. Costa, S. Costa, P. Arezes, J. M. López, and G. D. Arcas, “Improvement of performance in freezing of gait detection in parkinson’s disease using transformer networks and a single waist-worn triaxial accelerometer,” *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105482, 11 2022.

[9] S. A. Singh, T. G. Meitei, and S. Majumder, “Short pcg classification based on deep learning,” *Deep Learning Techniques for Biomedical and Health Informatics*, pp. 141–164, 1 2020.

[10] L. Goh, S. S. Paul, C. G. Canning, K. A. Ehgoetz Martens, J. Song, S. L. Campoy, and N. E. Allen, “The Ziegler Test is Reliable and Valid for Measuring Freezing of Gait in People With Parkinson Disease,” *Physical Therapy*, 09 2022. pzac122.

[11] S. Sveinbjornsdottir, “The clinical symptoms of parkinson’s disease,” *Journal of Neurochemistry*, vol. 139, pp. 318–324, 2016.

[12] T. Pringsheim, N. Jette, A. Frolkis, and T. D. Steeves, “The prevalence of parkinson’s disease: a systematic review and meta-analysis,” *Movement disorders*, vol. 29, no. 13, pp. 1583–1590, 2014.

[13] J. Nutt, B. Bloem, N. Giladi, M. Hallett, F. Horak, and A. Nieuwboer, “Freezing of gait: Moving forward on a mysterious clinical phenomenon,” *Lancet neurology*, vol. 10, pp. 734–44, 08 2011.

[14] A. H. Snijders, M. J. Nijkrake, M. Bakker, M. Munneke, C. Wind, and B. R. Bloem, "Clinimetrics of freezing of gait," *Movement Disorders*, vol. 23, pp. S468–S474, 1 2008.

[15] B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi, "Falls and freezing of gait in parkinson's disease: A review of two interconnected, episodic phenomena," *Movement Disorders*, vol. 19, pp. 871–884, 8 2004.

[16] K. Ziegler, F. Schroeteler, A. O. Ceballos-Baumann, and U. M. Fietzek, "A new rating instrument to assess festination and freezing gait in parkinsonian patients," *Movement disorders : official journal of the Movement Disorder Society*, vol. 25, pp. 1012–1018, 6 2010.

[17] A. E. Scully, D. Tan, B. I. R. de Oliveira, K. D. Hill, R. Clark, and Y. H. Pua, "Scoring festination and gait freezing in people with parkinson's: The freezing of gait severity tool-revised," *Physiotherapy Research International*, p. e2016, 2023.

[18] A. E. Scully, D. M. L. Tan, B. I. de Oliveira, K. D. Hill, R. Clark, and Y. H. Pua, "Validity and reliability of a new clinician-rated tool for freezing of gait severity," *Disability and Rehabilitation*, 2023.

[19] T. R. Morris, C. Cho, V. Dilda, J. M. Shine, S. L. Naismith, S. J. Lewis, and S. T. Moore, "A comparison of clinical and objective measures of freezing of gait in parkinson's disease," *Parkinsonism and Related Disorders*, vol. 18, pp. 572–577, 6 2012.

[20] G. Taghizadeh, S. M. Fereshtehnejad, P. Martinez-Martin, M. T. Joghataei, F. Mahdizadeh, S. Sabbaghi, S. Goudarzi, M. Meimandi, S. A. H. Habibi, and M. Mehdizadeh, "Clinimetrics of the freezing of gait questionnaire for parkinson disease during the "off" state," *Basic and Clinical Neuroscience*, vol. 12, p. 69, 1 2021.

[21] F. Hulzinga, A. Nieuwboer, B. W. Dijkstra, M. Mancini, C. Strouwen, B. R. Bloem, and P. Ginis, "The new freezing of gait questionnaire: Unsuitable as an outcome in clinical trials?," *Movement Disorders Clinical Practice*, vol. 7, p. 199, 2 2020.

[22] K. E. Martens, J. Shine, C. Walton, M. Georgiades, M. Gilat, J. Hall, A. Muller, J. Szeto, and S. Lewis, “Evidence for subtypes of freezing of gait in parkinson’s disease,” *Movement disorders : official journal of the Movement Disorder Society*, vol. 33, pp. 1174–1178, 1 2018.

[23] B. Thanvi and S. D. Treadwell, “Freezing of gait in older people: associated conditions, clinical aspects, assessment and treatment,” *Postgraduate Medical Journal*, vol. 86, pp. 472–477, 8 2010.

[24] C. K. Cui and S. J. Lewis, “Future therapeutic strategies for freezing of gait in parkinson’s disease,” *Frontiers in Human Neuroscience*, vol. 15, p. 741918, 11 2021.

[25] N. Giladi, “Medical treatment of freezing of gait,” *Movement Disorders*, vol. 23, pp. S482–S488, 2008.

[26] S. Pardoel, J. Kofman, J. Nantel, and E. D. Lemaire, “Wearable-sensor-based detection and prediction of freezing of gait in parkinson’s disease: A review,” *Sensors*, vol. 19, 2019.

[27] H. Cockx, J. Nonnekes, B. Bloem, R. van Wezel, I. Cameron, and Y. Wang, “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*, vol. 20, pp. 1–15, 12 2023.

[28] S. Rahman, H. J. Griffin, N. P. Quinn, and M. Jahanshahi, “The factors that induce or overcome freezing of gait in parkinson’s disease,” *Behavioural neurology*, vol. 19, pp. 127–136, 2008.

[29] M. Gilat, “How to annotate freezing of gait from video: A standardized method using open-source software,” *Journal of Parkinson’s Disease*, vol. 9, pp. 821–824, 1 2019.

[30] J. M. Shine, S. T. Moore, S. J. Bolitho, T. R. Morris, V. Dilda, S. L. Naismith, and S. J. Lewis, “Assessing the utility of freezing of gait questionnaires in parkinson’s disease,” *Parkinsonism & Related Disorders*, vol. 18, pp. 25–29, 1 2012.

[31] H. M. Parsons, "What happened at hawthorne?," *Science*, vol. 183, pp. 922–932, 3 1974.

[32] M. Mancini, V. V. Shah, S. Stuart, C. Curtze, F. B. Horak, D. Safarpour, and J. G. Nutt, "Measuring freezing of gait during daily-life: an open-source, wearable sensors approach," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, p. 1, 2021.

[33] B. L. Filkins, J. Y. Kim, B. Roberts, W. Armstrong, M. A. Miller, M. L. Hultner, A. P. Castillo, J. C. Ducom, E. J. Topol, and S. R. Steinhubl, "Privacy and security in the era of digital health: what should translational researchers know and do about it?," *American Journal of Translational Research*, vol. 8, p. 1560, 2016.

[34] A. Arami, A. Poulakakis-Daktylidis, Y. F. Tai, and E. Burdet, "Prediction of gait freezing in parkinsonian patients: A binary classification augmented with time series prediction.," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 27, pp. 1909–1919, 9 2019.

[35] S. Mazilu, M. Hardegger, Z. Zhu, D. Roggen, G. Tröster, M. Plotnik, and J. M. Hausdorff, "Online detection of freezing of gait with smartphones and machine learning techniques," *2012 6th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, PervasiveHealth 2012*, pp. 123–130, 2012.

[36] H. Zach, A. M. Janssen, A. H. Snijders, A. Delval, M. U. Ferraye, E. Auff, V. Weerdesteyn, B. R. Bloem, and J. Nonnekes, "Identifying freezing of gait in parkinson's disease during freezing provoking tasks using waist-mounted accelerometry," *Parkinsonism & Related Disorders*, vol. 21, pp. 1362–1366, 11 2015.

[37] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, 10 2007.

[38] M. T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of spearman's and kendall's correlation coefficients forassociation between two measured traits," *Animal Behaviour*, vol. 102, pp. 77–84, 4 2015.

[39] T. K. Kim, "Understanding one-way anova using conceptual figures," *Korean Journal of Anesthesiology*, vol. 70, p. 22, 2 2017.

[40] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.

[41] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," *Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019*, pp. 442–452, 10 2019.

[42] H. Cockx, E. Klaver, M. Tjepkema-Cloostermans, R. van Wezel, and J. Nonnekes, "The gray area of freezing of gait annotation: A guideline and open-source practical tool," *Movement Disorders Clinical Practice*, vol. 9, pp. 1099–1104, 11 2022.

[43] L. Borzì, L. Sigcha, D. Rodríguez-Martín, and G. Olmo, "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*, vol. 135, p. 102459, 1 2023.

[44] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, and B. Vanrumste, "Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops," *medRxiv*, p. 2023.05.05.23289387, 5 2023.

[45] T. Bikias, D. Iakovakis, S. Hadjidimitriou, V. Charisis, and L. J. Hadjileontiadis, "Deepfog: An imu-based detection of freezing of gait episodes in parkinson's disease patients via deep learning," *Frontiers in Robotics and AI*, vol. 8, 5 2021.

[46] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179–211, 3 1990.

[47] A. Howard, A. Salomon, E. Gazit, HCL-Jevster, J. Hausdorff, L. Kirsch, Maggie, P. Ginis, R. Holbrook, and Y. F. Karim, “Parkinson’s freezing of gait prediction,” 2023.

[48] B. Manor, M. Dagan, T. Herman, N. A. Gouskova, V. G. Vanderhorst, N. Giladi, T. G. Travison, A. Pascual-Leone, L. A. Lipsitz, and J. M. Hausdorff, “Multitarget transcranial electrical stimulation for freezing of gait: A randomized controlled trial,” *Movement disorders : official journal of the Movement Disorder Society*, vol. 36, pp. 2693–2698, 11 2021.

[49] A. Khan, N. Hammerla, S. Mellor, and T. Plötz, “Optimising sampling rates for accelerometer-based human activity recognition,” *Pattern Recognition Letters*, vol. 73, pp. 33–40, 4 2016.

[50] H. Zhou and H. Hu, “Human motion tracking for rehabilitation—a survey,” *Biomedical Signal Processing and Control*, vol. 3, pp. 1–18, 1 2008.

[51] L. Sigcha, N. Costa, I. Pavón, S. Costa, P. Arezes, J. M. López, and G. D. Arcas, “Deep learning approaches for detecting freezing of gait in parkinson’s disease patients through on-body acceleration sensors,” *Sensors (Basel, Switzerland)*, vol. 20, 4 2020.

[52] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[53] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 6 2017.

[55] S. Bouraya and A. Belangour, “Deep learning based neck models for object detection: A review and a benchmarking study,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, 2021.

[56] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[57] A. E. Patla, “Strategies for dynamic stability during adaptive human locomotion,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 22, pp. 48–52, 3 2003.

APPENDICES

Appendix A

Experiment Code

All code can be found on the following repository https://github.com/ahart97/fog_classification and is only available upon request.

Appendix B

Results Extended

B.1 Transformer Stock and Tuned Comparison

Below in Tables B.1 and B.2 are the comparison of the overall performance (i.e., on all Syndey dataset participants) for the tuned version of the transformer model, and the stock version with and without early stopping implemented.

Version	Early Stopping	Validation Loss	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Stock	FALSE	0.86	0.72	0.56	0.80	0.60	0.58
Stock	TRUE	0.33	0.72	0.51	0.83	0.62	0.56
Tuned	TRUE	0.42	0.74	0.83	0.69	0.58	0.69

Table B.1: Comparison of the version of transformer models in binary classification. Stock model with early stopping produces lowest (i.e., best) validation loss, while tuned model produces highest (i.e., best) F1-score.

Version	Early Stopping	Validation Loss	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Stock	FALSE	0.82	0.58	0.15	0.80	0.53	0.10
Stock	TRUE	0.41	0.59	0.09	0.86	0.03	0.05
Tuned	TRUE	0.85	0.59	0.47	0.66	0.42	0.42

Table B.2: Comparison of the version of transformer models in ternary classification. Stock model with early stopping produces lowest (i.e., best) validation loss, while tuned model produces highest (i.e., best) F1-score.

B.2 All Sensors Feature Engineering

When considering all sensors, the features selected for the binary and ternary FoG classification can be seen in Table B.3 and B.4 respectively.

Feature	**Sensor Types**	**Sensor Locations**	**Directions**
Root mean squared	Acc.	Sternum	AP
		Right Foot	AP, V
		Left Foot	AP
		Lumbar	AP, V
	Gyro.	Right Foot	AP, ML
		Left Foot	AP, V
		Lumbar	ML
Standard deviation	Acc.	Sternum	AP, ML, V
		Right Foot	AP, ML, V
		Left Foot	AP, ML, V
		Lumbar	AP, ML, V
	Gyro.	Sternum	AP, ML, V
		Right Foot	AP, ML, V
		Left Foot	AP, ML, V
		Lumbar	AP, ML, V
Kurtosis	Acc.	Sternum	ML
		Right Foot	AP
		Lumbar	AP
	Gyro.	Sternum	V
		Right Foot	AP, ML, V
		Left Foot	AP, V
Skewness	Acc.	Sternum	ML
		Right Foot	AP
		Left Foot	AP
Number of zero crossings	Acc.	Sternum	AP, ML, V
		Right Foot	AP, ML, V
		Left Foot	AP, V
		Lumbar	AP, ML

Table B.3: The 57 selected features based on ANOVA for binary freezing of gait classification when considering all sensors (AP - anterior-posterior, ML - medial-lateral, V - vertical).

Feature	**Sensor Types**	**Sensor Locations**	**Directions**
Root mean squared	Acc.	Sternum	AP
		Right Foot	AP, V
		Left Foot	AP
		Lumbar	AP, V
	Gyro.	Sternum	ML
		Right Foot	ML, AP, V
		Left Foot	AP
		Lumbar	ML, V
Standard deviation	Acc.	Sternum	AP, ML, V
		Right Foot	AP, ML, V
		Left Foot	AP, ML, V
		Lumbar	AP, ML, V
	Gyro.	Sternum	AP, ML, V
		Right Foot	AP, ML, V
		Left Foot	AP, ML, V
		Lumbar	AP, ML, V
Kurtosis	Acc.	Sternum	ML
		Lumbar	AP
	Gyro.	Sternum	V
		Right Foot	AP, ML, V
		Left Foot	AP, V
Skewness	Acc.	Sternum	ML
		Right Foot	AP
Number of dominant peaks	Acc.	Right Foot	ML
Number of zero crossings	Acc.	Sternum	AP, ML, V
		Right Foot	AP, V
		Left Foot	AP, V
		Lumbar	AP, ML

Table B.4: The 57 selected features based on ANOVA for ternary freezing of gait classification when considering all sensors (AP - anterior-posterior, ML - medial-lateral, V - vertical).

B.3 All Sensors Hyperparameter Tuning

The hyperparameters found when utilizing all sensors for the classical model can be seen in Table B.5.

Model	Parameter	Binary Tuned Value	Ternary Tuned Value
LR	Penalty	l1	l1
	Inverse of Regularization Strength	9.6	8.9
RF	Number of Estimators	116	50
	Criterion	entropy	entropy
	Max Features	log2	sqrt
SVM	Kernel	rbf	rbf
	Regularization	6.5	8.6
KNN	Number of Neighbours	5	10
	Weights	unifrom	distance
	Algorithm	brute	brute
	Leaf Size	40	100
	Power Parameter for the Minkowski Metric	1	1
ADT	Number of Estimators	66	150
	Criterion	entropy	log_loss
	Learning Rate	1.157895	2.868421
	Max Features	sqrt	auto
NN	Batch Size	256	64
	Activation	relu	relu
	Initial Learning Rate	0.01	0.01
	Number of Iterations with no Change	8	6
	Hidden Layer Size	90	80
	Solver	adam	adam

Table B.5: Optimal values found for the classical model hyperparameters during the tuning procedure for both binary and ternary classification with all sensors.

B.4 Additional Ternary Top Test Scores

The breakdown by the participant for the other three top models (ADT, SVM, and KNN) can be seen in Tables B.6 to B.8.

B.4.1 Lumbar Sensor

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.69	0.00	0.69	0.00	0.00	Mild (0.0%)	Moderate (22.31%)	Mild (0.0%)	Mild (8.46%)
85TL	0.96	0.00	0.97	0.00	0.00	Mild (0.0%)	Mild (1.53%)	Mild (0.76%)	Mild (1.53%)
76CA	0.75	0.00	0.80	0.00	0.00	Mild (2.49%)	Mild (2.85%)	Mild (3.56%)	Moderate (16.37%)
21DH	0.87	0.47	0.92	0.47	0.47	Mild (0.71%)	Mild (2.14%)	Mild (10.0%)	Mild (10.0%)
54EJ	0.78	0.04	0.87	0.11	0.06	Mild (2.29%)	Moderate (10.55%)	Mild (9.17%)	Mild (3.21%)
83OS	0.67	0.17	0.77	0.21	0.19	Mild (1.01%)	Mild (9.55%)	Moderate (16.58%)	Moderate (13.57%)
45PG	0.68	0.11	0.82	0.38	0.17	Mild (0.0%)	Moderate (16.55%)	Moderate (20.14%)	Mild (5.76%)
28FV	0.56	0.37	0.75	0.48	0.41	Moderate (32.5%)	Moderate (27.97%)	Moderate (16.92%)	Mild (6.7%)
97MU	0.62	0.49	0.88	0.86	0.62	Severe (64.57%)	Moderate (35.43%)	Mild (2.29%)	Mild (5.71%)
39KR	0.51	0.41	0.89	0.91	0.56	Severe (76.55%)	Moderate (34.23%)	Mild (1.35%)	Mild (8.89%)
Average	0.71	0.21	0.84	0.34	0.25	-	-	-	-
SD	0.13	0.20	0.08	0.32	0.23	-	-	-	-

Table B.6: Performance metrics for the AdaBoosted decision tree model (TS3) in ternary classification broken down by individual within the Sydney dataset.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.00	0.00	0.00	0.00	0.00	Mild (0.0%)	Severe (100.0%)	Mild (0.0%)	Mild (0.0%)
85TL	0.08	0.00	0.08	0.00	0.00	Mild (0.0%)	Severe (90.08%)	Mild (0.76%)	Mild (2.29%)
76CA	0.05	0.41	0.03	0.01	0.02	Mild (2.49%)	Severe (97.15%)	Mild (3.56%)	Mild (0.0%)
21DH	0.09	0.07	0.10	0.00	0.00	Mild (0.71%)	Severe (90.0%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.04	0.20	0.02	0.00	0.01	Mild (2.29%)	Severe (98.62%)	Mild (9.17%)	Mild (0.0%)
83OS	0.25	0.03	0.29	0.00	0.00	Mild (1.01%)	Severe (65.83%)	Moderate (16.58%)	Mild (0.5%)
45PG	0.19	0.00	0.24	0.00	0.00	Mild (0.0%)	Severe (66.19%)	Moderate (20.14%)	Mild (2.88%)
28FV	0.33	0.64	0.02	0.21	0.32	Moderate (32.5%)	Severe (97.65%)	Moderate (16.92%)	Mild (0.17%)
97MU	0.25	0.32	0.09	0.40	0.36	Severe (64.57%)	Severe (52.0%)	Mild (2.29%)	Mild (0.57%)
39KR	0.68	0.86	0.02	0.73	0.79	Severe (76.55%)	Severe (90.03%)	Mild (1.35%)	Mild (0.27%)
Average	0.19	0.25	0.09	0.14	0.15	-	-	-	-
SD	0.19	0.29	0.10	0.24	0.25	-	-	-	-

Table B.7: Performance metrics for the support vector machine model (TM4) in ternary classification broken down by individual within the Sydney dataset.

B.4.2 All Sensors

The breakdown by the participant for the other top four models (NN, SVM, RF, and KNN) when considering all sensors instead of just the lumbar accelerometer can be seen in Tables B.9 to B.12.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.79	0.00	0.79	0.00	0.00	Mild (0.0%)	Moderate (20.0%)	Mild (0.0%)	Mild (0.77%)
85TL	0.98	0.00	0.99	0.00	0.00	Mild (0.0%)	Mild (0.76%)	Mild (0.76%)	Mild (0.0%)
76CA	0.66	0.35	0.68	0.03	0.06	Mild (2.49%)	Moderate (28.83%)	Mild (3.56%)	Mild (4.98%)
21DH	0.84	0.07	0.93	0.00	0.01	Mild (0.71%)	Moderate (10.71%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.83	0.00	0.94	0.00	0.00	Mild (2.29%)	Mild (5.96%)	Mild (9.17%)	Mild (0.92%)
83OS	0.53	0.11	0.62	0.21	0.09	Mild (1.01%)	Moderate (39.2%)	Moderate (16.58%)	Mild (4.52%)
45PG	0.60	0.00	0.75	0.00	0.00	Mild (0.0%)	Moderate (28.78%)	Moderate (20.14%)	Mild (0.0%)
28FV	0.57	0.23	0.90	0.50	0.29	Moderate (32.5%)	Moderate (18.93%)	Moderate (16.92%)	Mild (0.5%)
97MU	0.46	0.25	0.88	0.81	0.36	Severe (64.57%)	Moderate (17.71%)	Mild (2.29%)	Moderate (15.43%)
39KR	0.40	0.25	0.93	0.91	0.38	Severe (76.55%)	Moderate (20.49%)	Mild (1.35%)	Mild (4.31%)
Average	0.67	0.13	0.84	0.25	0.12	-	-	-	-
SD	0.18	0.13	0.12	0.34	0.15	-	-	-	-

Table B.8: Performance metrics for the k-nearest neighbours model (TS5) in ternary classification broken down by individual within the Sydney dataset.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.96	0.00	0.96	0.00	0.00	Mild (0.0%)	Mild (3.85%)	Mild (0.0%)	Mild (0.0%)
85TL	0.97	0.00	0.98	0.00	0.00	Mild (0.0%)	Mild (2.29%)	Mild (0.76%)	Mild (0.0%)
76CA	0.86	0.41	0.89	0.07	0.13	Mild (2.49%)	Moderate (13.88%)	Mild (3.56%)	Mild (0.0%)
21DH	0.89	0.00	1.00	0.00	0.00	Mild (0.71%)	Mild (0.0%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.89	0.00	1.00	0.00	0.00	Mild (2.29%)	Mild (0.0%)	Mild (9.17%)	Mild (0.0%)
83OS	0.79	0.00	0.96	0.00	0.00	Mild (1.01%)	Mild (4.52%)	Moderate (16.58%)	Mild (0.0%)
45PG	0.80	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Moderate (20.14%)	Mild (0.0%)
28FV	0.61	0.58	0.63	0.45	0.49	Moderate (32.5%)	Moderate (49.41%)	Moderate (16.92%)	Mild (9.05%)
97MU	0.34	0.01	1.00	0.97	0.02	Severe (64.57%)	Mild (0.57%)	Mild (2.29%)	Mild (0.0%)
39KR	0.90	0.94	0.76	0.91	0.92	Severe (76.55%)	Severe (79.25%)	Mild (1.35%)	Mild (0.0%)
Average	0.80	0.19	0.92	0.24	0.16	-	-	-	-
SD	0.18	0.32	0.12	0.37	0.29	-	-	-	-

Table B.9: Performance metrics for the neural network model in ternary classification with all sensors broken down by individual within the Sydney dataset.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	0.82	0.00	0.82	0.00	0.00	Mild (0.0%)	Mild (2.31%)	Mild (0.0%)	Moderate (15.38%)
85TL	0.49	0.00	0.49	0.00	0.00	Mild (0.0%)	Mild (1.53%)	Mild (0.76%)	Moderate (48.85%)
76CA	0.37	0.71	0.35	0.11	0.18	Mild (2.49%)	Mild (9.96%)	Mild (3.56%)	Severe (56.94%)
21DH	0.85	0.53	0.89	0.32	0.40	Mild (0.71%)	Mild (0.0%)	Mild (10.0%)	Moderate (16.43%)
54EJ	0.61	0.60	0.61	0.13	0.21	Mild (2.29%)	Mild (0.0%)	Mild (9.17%)	Moderate (43.12%)
83OS	0.62	0.77	0.59	0.27	0.40	Mild (1.01%)	Mild (0.5%)	Moderate (16.58%)	Moderate (47.24%)
45PG	0.79	0.50	0.86	0.48	0.49	Mild (0.0%)	Mild (0.0%)	Moderate (20.14%)	Moderate (20.86%)
28FV	0.42	0.33	0.50	0.40	0.30	Moderate (32.5%)	Moderate (14.41%)	Moderate (16.92%)	Moderate (44.22%)
97MU	0.31	0.03	0.88	0.00	0.00	Severe (64.57%)	Mild (0.0%)	Mild (2.29%)	Severe (60.0%)
39KR	0.91	0.97	0.67	0.88	0.93	Severe (76.55%)	Severe (84.37%)	Mild (1.35%)	Mild (0.54%)
Average	0.62	0.44	0.67	0.26	0.29	-	-	-	-
SD	0.21	0.33	0.18	0.26	0.27	-	-	-	-

Table B.10: Performance metrics for the support vector machine model in ternary classification with all sensors broken down by individual within the Sydney dataset.

Participant	Accuracy	Sensitivity	Specificity	Precision	F1-Score	True Akin. Severity	Predicted Akin. Severity	True Kin. Severity	Predicted Kin. Severity
93QN	1.00	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Mild (0.0%)	Mild (0.0%)
85TL	0.99	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Mild (0.76%)	Mild (0.0%)
76CA	0.86	0.24	0.91	0.05	0.09	Mild (2.49%)	Moderate (11.03%)	Mild (3.56%)	Mild (0.71%)
21DH	0.89	0.00	1.00	0.00	0.00	Mild (0.71%)	Mild (0.0%)	Mild (10.0%)	Mild (0.0%)
54EJ	0.89	0.00	1.00	0.00	0.00	Mild (2.29%)	Mild (0.0%)	Mild (9.17%)	Mild (0.0%)
83OS	0.81	0.00	0.99	0.00	0.00	Mild (1.01%)	Mild (0.5%)	Moderate (16.58%)	Mild (0.5%)
45PG	0.80	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Moderate (20.14%)	Mild (0.0%)
28FV	0.62	0.31	0.93	0.42	0.35	Moderate (32.5%)	Moderate (23.45%)	Moderate (16.92%)	Mild (0.0%)
97MU	0.37	0.06	0.98	0.85	0.11	Severe (64.57%)	Mild (4.57%)	Mild (2.29%)	Mild (2.86%)
39KR	0.75	0.72	0.84	0.92	0.81	Severe (76.55%)	Severe (60.11%)	Mild (1.35%)	Mild (0.0%)
Average	0.80	0.13	0.96	0.22	0.14	-	-	-	-
SD	0.18	0.22	0.05	0.35	0.25	-	-	-	-

Table B.11: Performance metrics for the random forest model in ternary classification with all sensors broken down by individual within the Sydney dataset.

Participant	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-Score**	**True Akin. Severity**	**Predicted Akin. Severity**	**True Kin. Severity**	**Predicted Kin. Severity**
93QN	0.95	0.00	0.95	0.00	0.00	Mild (0.0%)	Mild (3.85%)	Mild (0.0%)	Mild (0.77%)
85TL	0.98	0.00	0.98	0.00	0.00	Mild (0.0%)	Mild (1.53%)	Mild (0.76%)	Mild (0.0%)
76CA	0.70	0.29	0.73	0.05	0.09	Mild (2.49%)	Moderate (14.59%)	Mild (3.56%)	Moderate (13.88%)
21DH	0.89	0.00	0.99	0.00	0.00	Mild (0.71%)	Mild (1.43%)	Mild (10.0%)	Mild (0.71%)
54EJ	0.85	0.00	0.96	0.00	0.00	Mild (2.29%)	Mild (5.96%)	Mild (9.17%)	Mild (0.0%)
83OS	0.79	0.00	0.96	0.00	0.00	Mild (1.01%)	Mild (4.02%)	Moderate (16.58%)	Mild (0.0%)
45PG	0.80	0.00	1.00	0.00	0.00	Mild (0.0%)	Mild (0.0%)	Moderate (20.14%)	Mild (0.0%)
28FV	0.52	0.19	0.85	0.34	0.24	Moderate (32.5%)	Moderate (19.6%)	Moderate (16.92%)	Mild (3.02%)
97MU	0.45	0.20	0.95	0.89	0.32	Severe (64.57%)	Moderate (14.29%)	Mild (2.29%)	Moderate (13.71%)
39KR	0.75	0.75	0.74	0.90	0.81	Severe (76.55%)	Severe (63.88%)	Mild (1.35%)	Mild (8.09%)
Average	0.77	0.14	0.91	0.22	0.15	-	-	-	-
SD	0.16	0.23	0.10	0.35	0.25	-	-	-	-

Table B.12: Performance metrics for the k-nearest neighbours model in ternary classification with all sensors broken down by individual within the Sydney dataset.

B.5 Strip Charts

The strip charts for the top two models (LR and the transformer) for the remaining participants can be seen in Figures B.1 to B.7.

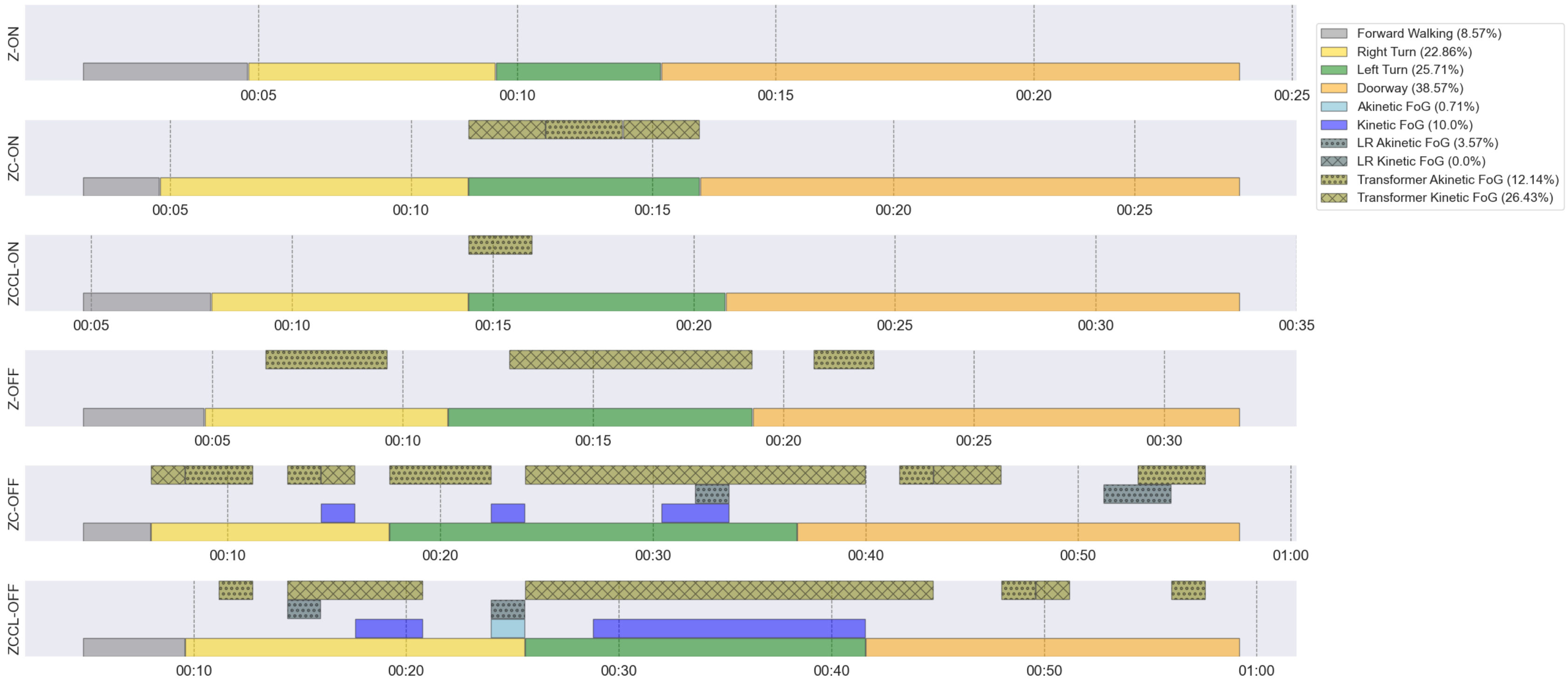


Figure B.1: Strip chart of all true and predicted labels for each of participant 21DH's trials.

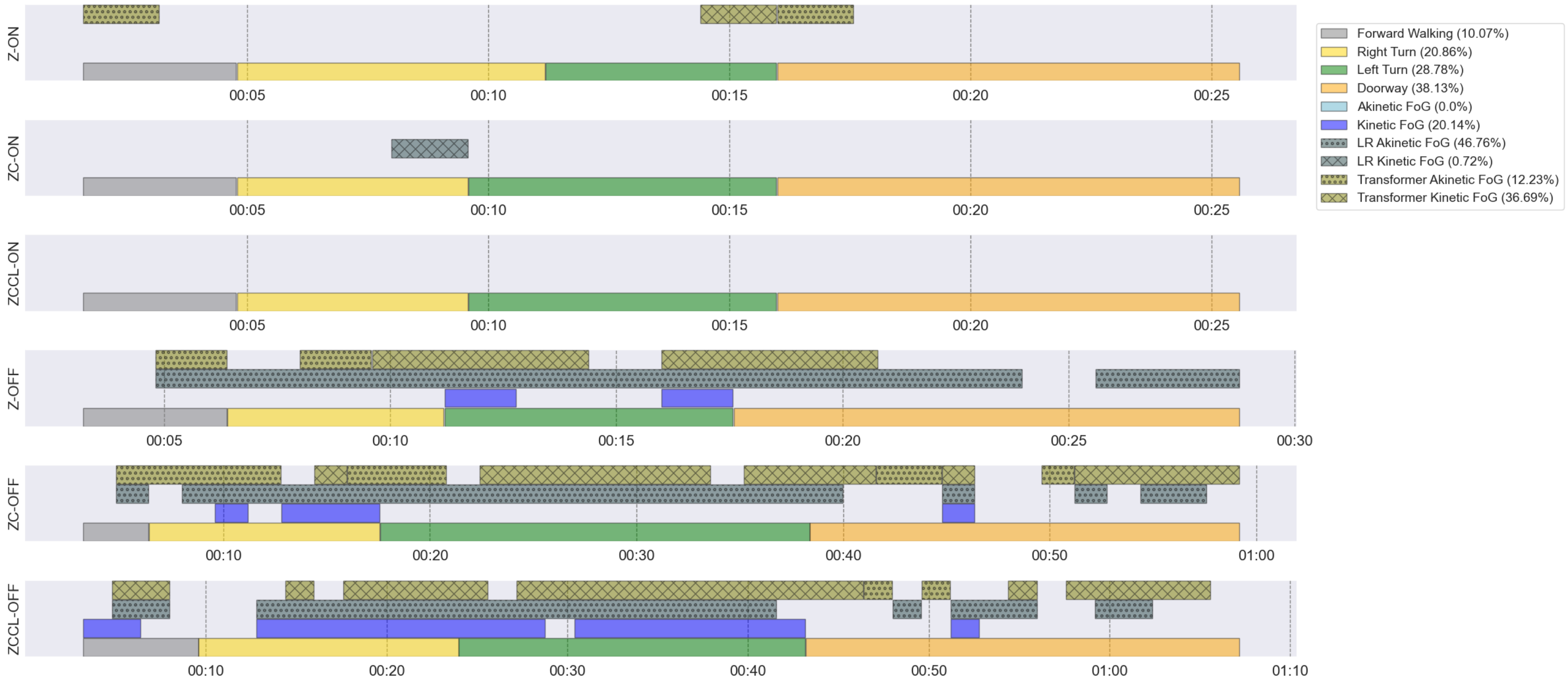


Figure B.2: Strip chart of all true and predicted labels for each of participant 45PG's trials.

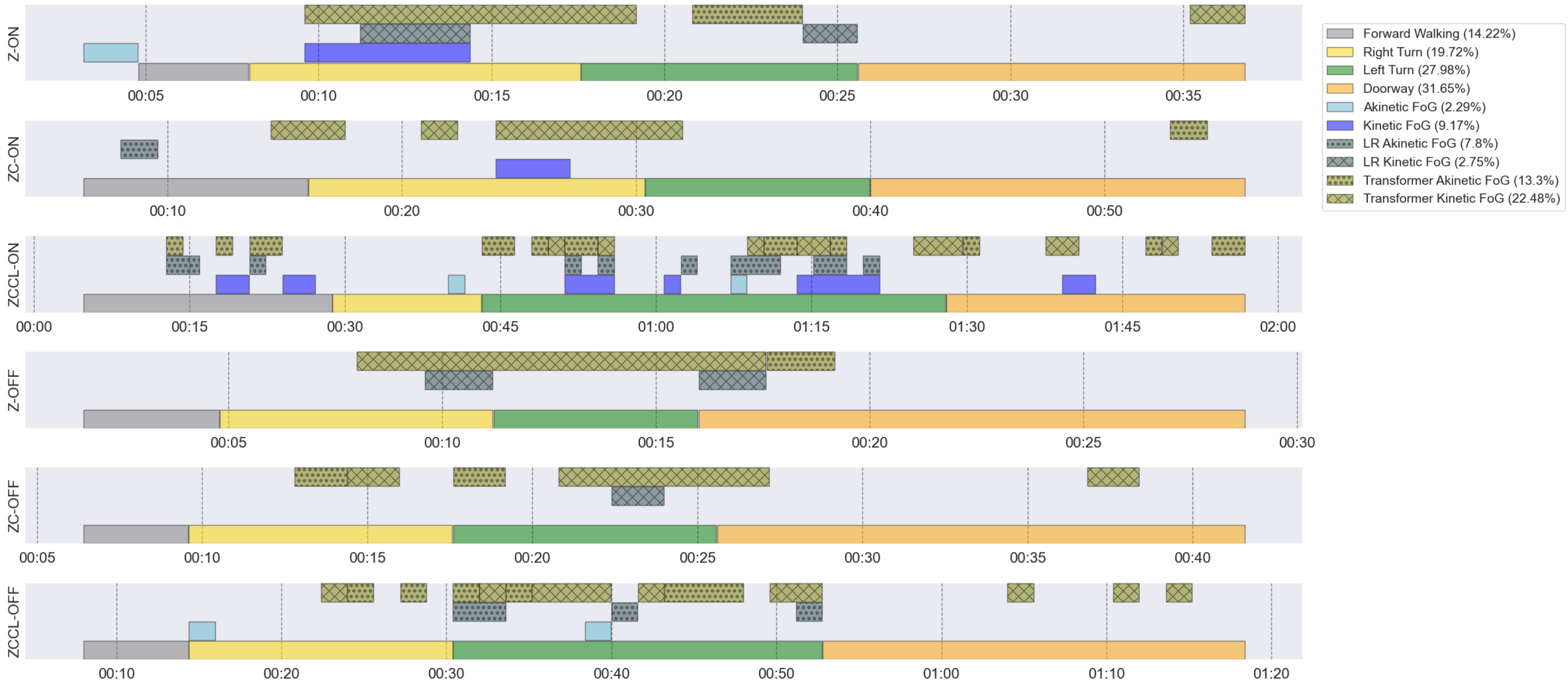


Figure B.3: Strip chart of all true and predicted labels for each of participant 54EJ's trials.

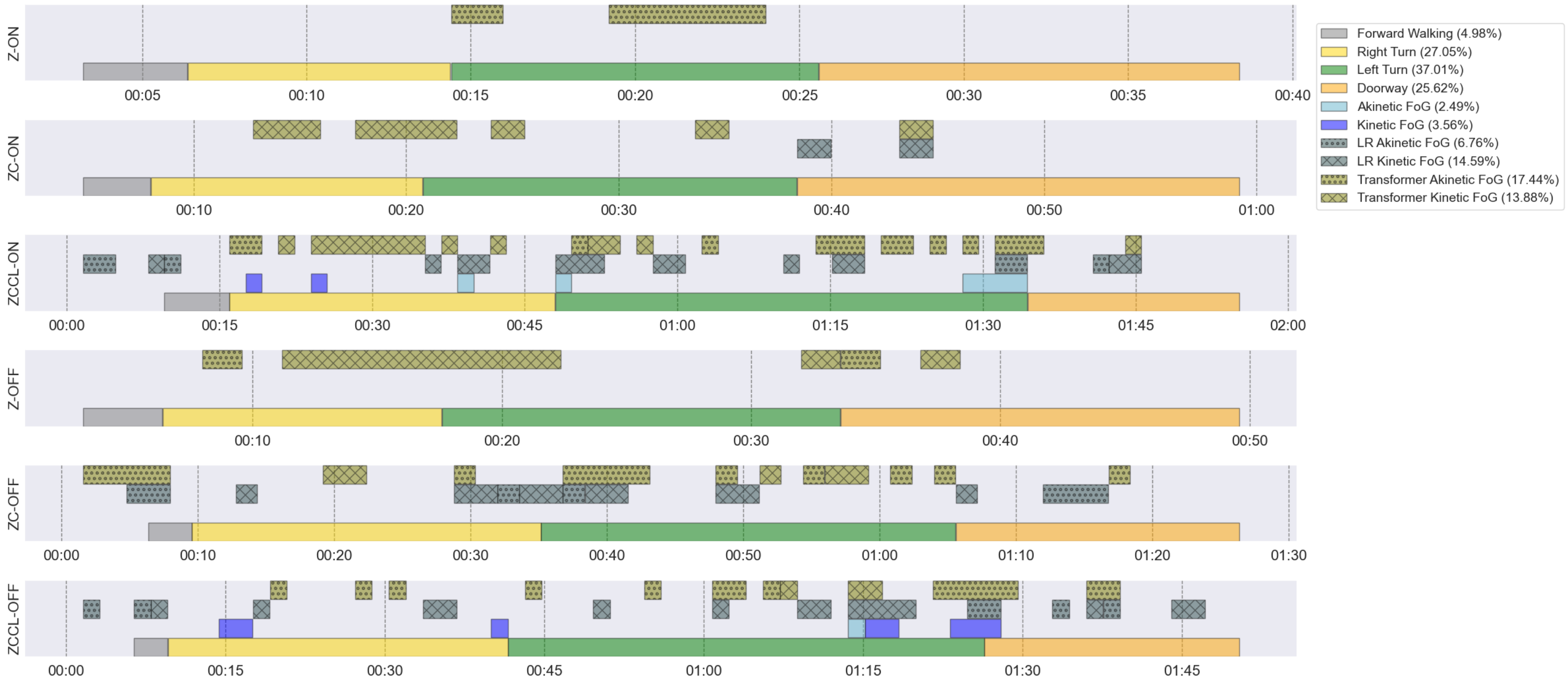


Figure B.4: Strip chart of all true and predicted labels for each of participant 76CA's trials.

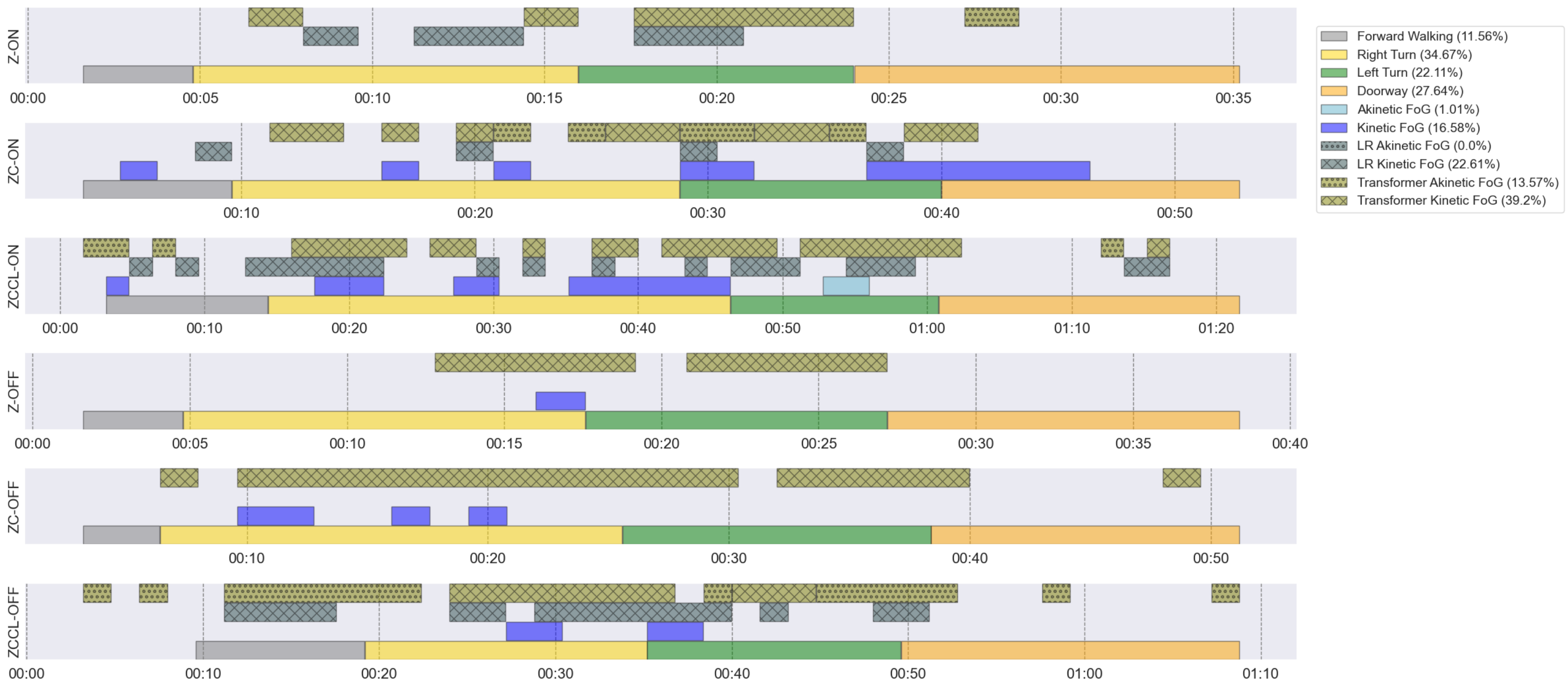


Figure B.5: Strip chart of all true and predicted labels for each of participant 83OS's trials.

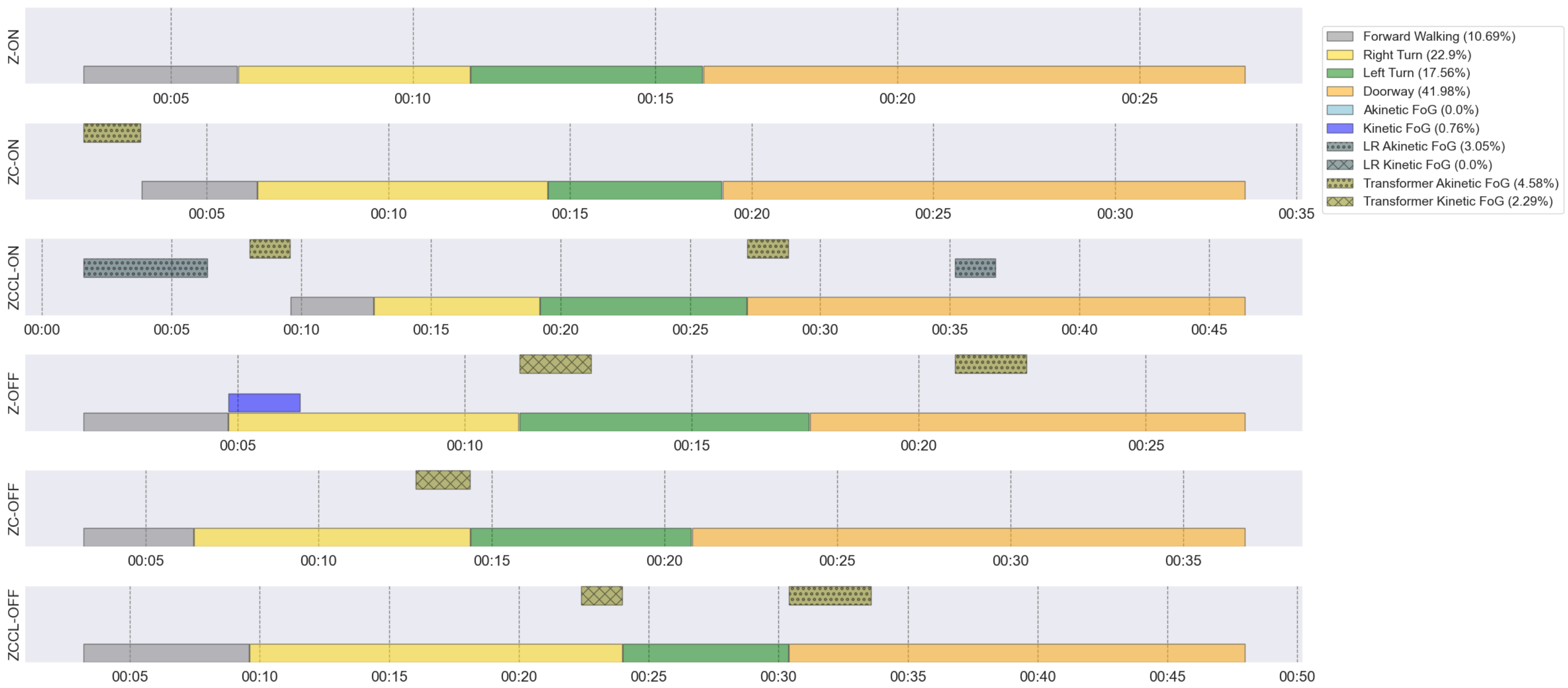


Figure B.6: Strip chart of all true and predicted labels for each of participant 85TL's trials.

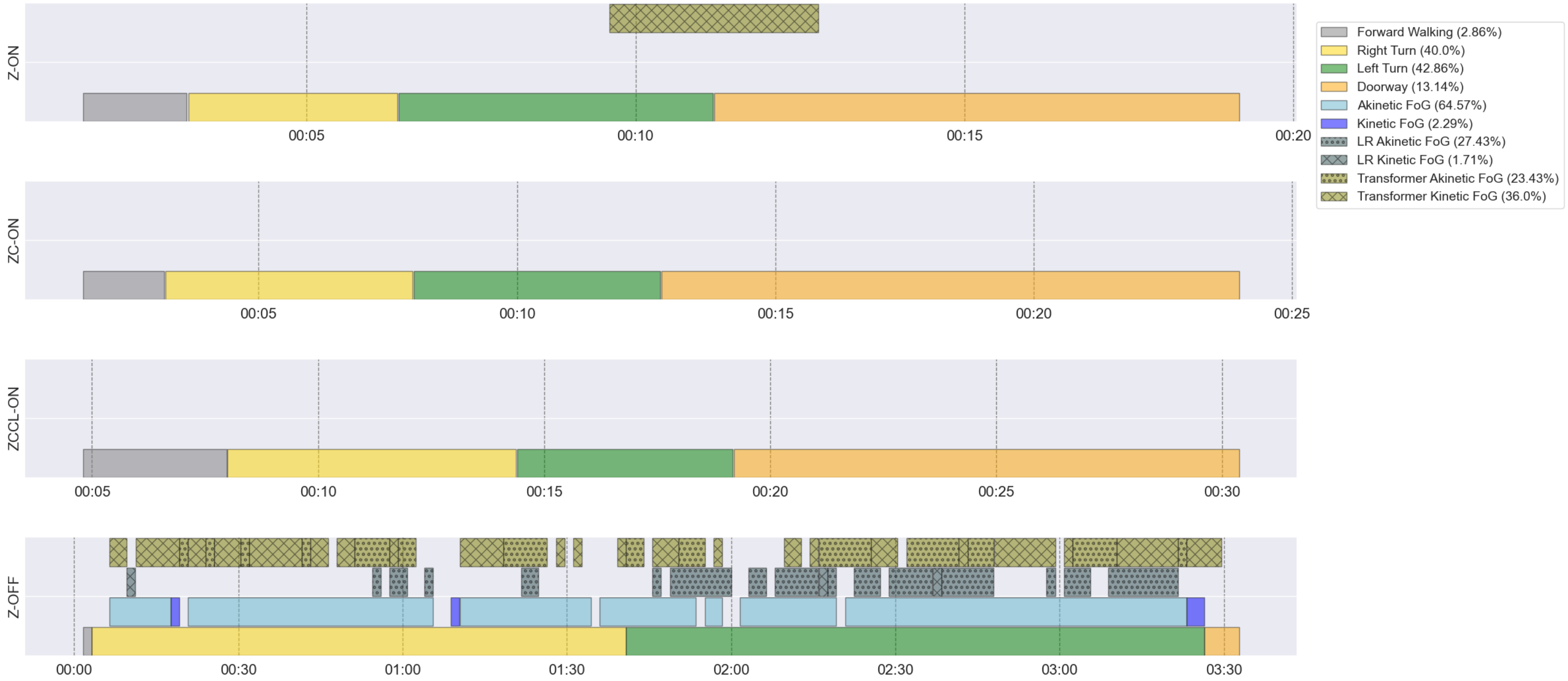


Figure B.7: Strip chart of all true and predicted labels for each of participant 97MU's trials.