Machine Learning Methods for Multiple Sclerosis Detection Based on Raw Data from an Instrumented Walkway

By Wenting Hu

A Thesis submitted

to the School of Graduate Studies in partial fulfillment of the

requirements for the degree of

Master of Science, Department of Computer Science, Faculty of Science
Memorial University of Newfoundland

March 2022
St. John’s Newfoundland and Labrador

Abstract

Background: Multiple sclerosis (MS) patients may experience varying gait disorders. In MS disease evaluation and diagnosis, instrumented walkways using embedded pressure sensors are widely used to provide information regarding gait disturbances. The information is delivered as predefined parameters, which may obscure salient features and patterns in the raw sensor data. This thesis applied machine learning techniques to raw walkway data to distinguish MS patients from healthy controls while further distinguishing the impairment levels of MS patients.

Methods: New features were constructed to supplement the standard parameters. A severity level was determined using patients' ratings of the severity of their gait problems on the MS Impact Scale-29. Two experiments were conducted. The first experiment focused on discerning healthy controls from MS patients. The second experiment attempted to classify patients with different impairment levels.

Results: The MS vs. Healthy experiment achieved a good baseline accuracy of 81% using the standard feature set and received a 7% improvement using the augmented set. The mild MS vs. moderate MS experiment achieved an accuracy of 76% using the standard set, which was further improved by 2% using the augmented set.

Conclusion: These experiments demonstrate that the newly generated features improve the machine learning model results with excellent accuracy and precision.

Acknowledgments

The experiments in this thesis were under the supervision of Dr. Xianta Jiang and Dr. Michelle Ploughman. I would like to thank them for their consistent guidance and invaluable support sincerely. I would also like to thank Owen Combden for his kind help in building the first experiment with me. Also, I acknowledge the Recovery & Performance Laboratory members for their help with data collection and for taking the time to review my thesis. I also acknowledge the help from the Ubiquitous Computing and Machine Learning Research Lab members who provide technical support. Finally, I would like to thank my parents for their trust and love and my boyfriend Zhongtian, whose great encouragement has supported me throughout this journey.

Contents

List of Tables

List of Figures

List of Abbreviations

ADASYN	Adaptive synthetic sampling approach
ANOVA	Analysis of Variance F-Test statistics
AUC	Area Under Curves
AUPRC	Area Underneath the Precision-Recall
AUROC	Area Underneath the Receiver Operating Characteristic
BOS	Base of Support
DT	Dual-task Walking Test
EDSS	Expanded Disability Status Scale
FN	False Negative
FP	False Positive
HITMS	Health Innovation Team in MS
KNN	K Nearest Neighbors
LOP	Line of Progression
LR	Logistic Regression
MS	Multiple Sclerosis
MSIS	Multiple Sclerosis Impact Scale
PR	Precision-Recall
PRC	Precision-Recall Curve
RFECV	Recursive Feature Elimination with Cross-Validation
ROC	Receiver Operating Characteristic
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling TEchnique
SS	self-selected speed walking test
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
W-AMS	Wearable Activity Monitoring System
XGB	XGBoost

1 Introduction

Multiple sclerosis (MS) is a common inflammatory neurodegenerative central nervous system disease [1] with a prevalence of 25 per 100,000 people in Canada and 35.9 per 100,000 people worldwide [2]. MS symptoms, which include slower information processing, walking impairments, and feelings of mental fatigue, profoundly impact the patient's quality of life [3]. People with MS experience gait disorders such as spasticity, leg weakness, foot drop, and ataxia, which interfere with their daily tasks [4]–[6]. These disorders may differ from person-to-person, likely due to the unique patterns of central nervous system lesions and neural reorganization [1], [7]. To examine gait disorders, therapists use visual inspection and clinical evaluation.

Expanded Disability Status Scale (EDSS) is used by neurologists to assess the degree of disability among MS patients. The EDSS scores range from 0 to 10; higher scores indicate more disability. Although EDSS is widely used in the MS field, it is relatively insensitive to gait changes and is insufficient in characterizing patients with mild disabilities [8].

Patients may identify symptom changes before neurologists detect declines or improvements. One MS-related patient-reported evaluation is the Multiple Sclerosis Impact Scale (MSIS-29). It includes 29 questions: 20 items about physical function and 9 describing cognitive and mental health problems. Ranging from 1 to 5, the higher the score, the more significant the impact of that symptom on everyday life is. Sample MSIS-29 questions include assessments of patient's ability to use fine motor skills or questions on balance and coordination, such as using their hands or walking difficulties [9]. For patients in the early stages of the disease, MSIS-29 helps the clinician track changes in their mental and physical health. Both

EDSS and MSIS-29 emphasize the importance of walking, suggesting that gait detection and analysis are vital neurological benchmarks that can guide treatment decisions [10].

Most studies examining gait changes in MS focus on reductionist methods, reporting output variables such as walking velocity or distance walked. Newer technologies and analytical techniques provide expanded opportunities to identify unique gait patterns within and between individuals. Such innovations can help detect early changes in direct rehabilitation interventions to improve walking [11], [12]. For example, using image-processing techniques [13] and wearable sensors, users can create movement-related features in a three-dimensional coordinate system [14], [15]. However, these methods require specialized equipment that is not readily available to clinicians. Data-driven techniques, such as machine learning classification, enable the analysis of specific gait features and their relationship with the disease. For instance, Chen et al. employed machine learning to identify gait variables extracted from walking and jumping tests in order to classify patients with mild cognitive impairment [16]. Data gathered from vertical ground reaction force sensors were provided to machine learning algorithms to successfully detect the early signs of Parkinson's disease [17]. In the field of MS, one study used machine learning techniques to detect which gait parameters were most sensitive to subtle gait changes [18]. However, the study used the predetermined and limited gait variables available in conventional proprietary software. The instrumented walkway is an assistive tool increasingly used to sensitively map the spatiotemporal profile of gait. The instrumented walkway consists of a dense matrix of embedded sensors to capture temporal, spatial, and force-related gait data from footsteps. Depending on the participant characteristics and the length of

the mat, one pass across the walkway captures 4 to 10 footsteps, thus generating thousands of data points. Walkway systems use secondary software packages to transform raw sensor data into a standard set of output variables that are useful for clinicians, such as walking speed and step length [19], [20]. By interrogating the raw data, subtle changes in gait patterns, which are signs of disease worsening/improvement, can be revealed [21]. Machine learning can be used to build models that reflect the relationships between gait variables and disease. Creating novel gait variables from raw walkway data may increase model accuracy and precision, pinpointing gait characteristics that require clinical attention. This may result in tailored, individualized, and effective rehabilitation strategies for gait training.

This thesis aims to determine whether newly designed gait features, calculated directly from raw walkway sensor data, can be used to increase predictive performance when classifying patients from healthy controls and to determine the severity of MS impairment. Two classification experiments were conducted. One was to classify MS patients from healthy controls. The other classified patients with lower (mild severity) or higher (moderate severity) MSIS scores. Both classification processes used machine learning technology combined with raw data obtained from an electronic walkway (Protokinetics Havertown PA). For both experiments, two series of analyses were performed to enable classification with a standard set of features as well as an expanded set of features that included several new or underutilized parameters derived from raw data.

It is hypothesized that machine learning models can effectively distinguish MS patients from healthy controls and patients with mild and moderate MS using only standard features.

These new or unutilized features would further improve detection accuracy. Such methodologies could be critical in automatically and accurately detecting subtle gait changes indicative of improvement or worsening neurological impairment.

1.1 Thesis Outline

In Chapter 2 of this thesis, previous and related work on how researchers detect and classify MS patients with various gait disorders are discussed, and the essential concepts used in this thesis are presented. In Chapter 3, the methodology is described, including two classification experiments. One is to classify MS patients from healthy controls. The other is to classify patients based on lower or higher MSIS scores. Both classification processes use machine learning technology combined with gait measurements calculated from raw walkway sensor data. Two experiments were performed for each classification experiment, with the first attempting to classify with a standard set of features. The second uses an expanded set of features that include standard features as well as several new or underutilized parameters derived from the raw data. In Chapter 4, the results of the two experiments are explained. Chapter 5 summarizes the findings of this thesis and discusses possible future improvements.

1.2 Contributions

To the best of our knowledge, this study is the first attempt to distinguish MS patients from healthy controls and to classify mild to moderate disease severity using machine learning on raw walkway sensor data. The first experiment, classifying MS patients from healthy controls [22], has been published. I have mainly contributed to the model training, testing, and validation

process in machine learning and manuscript preparation. In both experiments, raw sensor data provided opportunities to create new features. This study proved that new features improved the model performance. The use of MSIS-29 instead of EDSS as the standard for distinguishing the level of gait disorder is also new in classification experiments, which also raises the possibility of detecting self-reported problems.

2 Literature Review

The gait of MS patients is somewhat different from that of healthy people due to the disease, and the gait disorder usually deteriorates over time as the disease progresses. Clinicians and therapists either visually inspect gait or calculate walking speed to determine the severity of a patient's condition. Machine learning is a valuable tool for dealing with the enormous amount of data generated by a typical set of walking tools and examining the relationship between the features and the disease.

2.1 Gait & Gait Detection

The movement of gait requires the cooperation of the entire nervous system. The nervous system controls walking stability and standing posture to prevent people from falling. The signal generated from the brainstem and spinal cord is passed to the motor neurons to power the movement of the lower limbs. Signals from the visual system, basal ganglia, and cerebellum help coordinate the muscles of the limbs. Unusual changes in the nervous system can lead to gait disorders [23].

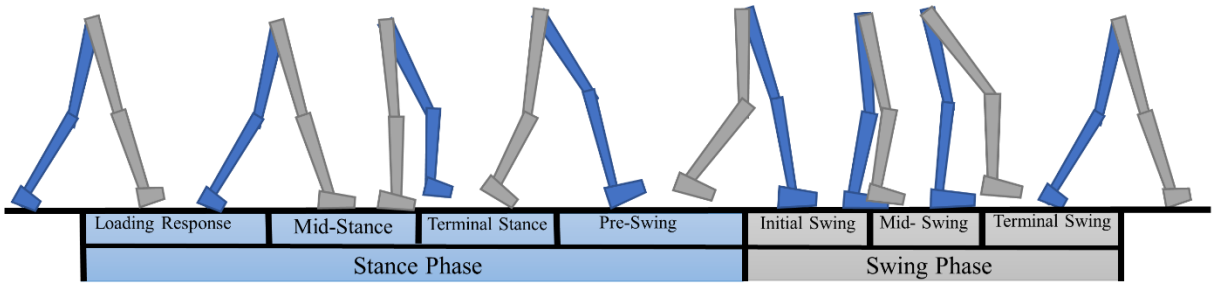

Figure 2.1 Gait cycle and gait phases. Stance phase contains loading response, mid-stance, terminal stance and pre-swing events. Swing phase contains initial swing, mid-swing and terminal swing events [11].

The gait contains continuous cycles, and each cycle can be divided into two phases: swing and stance phases (Figure 2.1). Each phase includes seven or eight events that switch from one event to another according to different definitions [24]–[26]. In the stance phase, the leading limb initiates the contact with the ground, and it consists of four events: loading response, mid-stance, terminal-stance and pre-swing. The heel strike of the leading limb occurs at the beginning of the loading response. The foot flat of the leading limb marks the beginning of mid-stance, while the heel off of the other limb in mid-stance and terminal-stance, and finally, the toe-off of the leading limb is the end action of the pre-swing. The stance phase ends, and the swing phase begins. The swing phase consists of three events: initial swing, mid-swing, and terminal swing, starting with the toe-off of the leading limb and lasting to the end of the toe-off of the other limb. Finally, the heel strike of the leading limb initiates a new gait cycle [11].

Detecting impairments in gait provides an indication of changes to the nervous system. Visual systems, wearable sensors, and electronic walkways are the most common gait detection tools.

2.1.1 Visual Recording of Gait

Video recording permits clinicians and therapists to replay how participants walk and to compare the walking differences over time. Researchers Decavel and Sagawa, in 2019, recorded how their patients walked from the front, back, and side in all walking tests. The recording allowed several evaluators to simultaneously evaluate gait impairment in patients [27]. The recording also provides the exact timing of foot contact and foot-off to differentiate

gait cycles. In another study, researchers designed a detection technique to automatically classify foot contact and foot-off by using a multilayer neural network. They used a motion capture system and force plates to collect walking data. The researchers then extracted features from the data by visually labeling the gait events. The trained model accurately detected foot contact and foot-off using the visual data [28]. 3-D gait analysis is another method for assessing gait [29]. This method reconstructs movement detected from reflective markers to produce 3-D images of gait. Thus, kinematic, kinetic, and spatiotemporal parameters are calculated for subtle gait impairments and abnormal posture detection. However, such visual data collection methods require expensive equipment and professional operators that are not commonly available in clinical settings.

2.1.2 Wearable Sensors

Wearable sensors allow clinicians and medical researchers to monitor participants' daily activities outside the laboratory. Meyer et al., in 2021, used wearable sensors and deep learning to detect fall risk for MS patients. They used wearable sensors to record the participant's one-minute trials in a hallway. Combined with the use of deep learning, they were able to successfully identify MS patients who had recently fallen and achieved an AUC of 88% [30]. Han et al., in 2007, developed a wearable activity monitoring system (W-AMS) to record the acceleration of both ankles in patients with Parkinson's disease to detect gait events. This system can synchronize patients' ankle acceleration with a video and foot pressure system. Five healthy controls and five Parkinson's patients were engaged in this experiment. Participants

wore W-AMS on both ankles to measure the three-dimensional acceleration. The detection accuracy was 93% for healthy people and 94% for Parkinson's patients [31].

A recent study by Zhao et al. aimed to design a system to detect gait events. This study detected gait using two foot-mounted inertial sensors, one on each heel. Six gait events, heel strike, foot flat, mid-stance, heel off, toe-off and mid-swing, were analyzed. The researchers used a neural network to extract features from raw data and provided the features as input into the hidden Markov model. This research showed that the hidden Markov model could classify different gait events using the sensor data [32]. Although wearable sensors were proven to be helpful in these studies, the accuracy of the wearable sensors varied depending on the sensor type or attached location [33], [34]. The unwillingness of participants [35] and sensor running battery life may also be limitations of wearable sensors.

2.1.3 Walkways

Walkways provide researchers with various information, for example, the walking distance, gait speed, and pressure distribution of the foot while people walk through it. Researchers have built and developed different walkways to adapt to different research purposes. For example, interactive walkways combined with external sensors provide clinicians or researchers with 3D full-body motion information and enable more task-specific walking and comprehensive gait evaluation by projecting visual context on the walkway [36], [37].

Figure 2.2 GAITRite electronic walkway is embedded with thousands of sensors and attached to the floor.

Electronic walkway (Figure 2.2) is a relatively traditional walkway. It contains a dense matrix of embedded sensors to capture temporal, spatial, and force-related gait data from footsteps, which is now commonly used among patients with MS. While participants walk across the walkway, the embedded sensors are activated when pressure is detected, and the walking data are recorded. Depending on the participant in the study and the length of the walkway mat, the walkway captures four to ten footsteps for one pass across and generates thousands of data points. A secondary software package transforms the raw sensor data into a standard set of output variables, such as walking speed, step length, and cadence, which are useful for clinicians [19], [20]. The predetermined features provided by the software are not only gait measurements but also composite numerical values that represent the gait impairment situation. These numerical values were calculated according to spatial and temporal data and

served as features for different studies [19]. In 2017 and 2018, Kirkland et al. used the walkway to measure hopping variables to detect mild impairments in patients. The footprint location and degree of pressure provided by the walkway made it possible to calculate hop length, hop time, and velocity, which correlated well with the participant's anticipatory motor control ability [38], [39]. Leone et al., in 2018, designed a study to examine the effect of rehabilitation in MS patients using gait parameters from the walkway system. The results showed that the gait parameters acquired from the walkway proved that after rehabilitation, the patients' walking condition improved [40]. The validity of using an electronic walkway to capture gait variables was demonstrated by the studies mentioned above. However, specific features can be created from raw sensor data using other machine learning techniques, and these newly created features may aim at specific gait disorders to further increase the accuracy of the studies.

2.2 Overview of Multiple Sclerosis

MS is a common inflammatory neurodegenerative central nervous system disease [1]. Ninety thousand Canadians live with MS, making Canada one of the countries with a high prevalence of MS. Approximately 2.8 million people live with MS worldwide (2020). This lifelong disease is diagnosed between the ages of 20 and 50 years [2].

2.2.1 The Effect of MS

MS affects the quality of patients' lives in both cognitive and physical aspects. MS may affect patients' ability to walk, long-term memory, and efficiency of information processing as well as fatigue and depression [41].

MS patients walk differently from healthy people due to limb weakness, sensory loss, and foot drop. These gait disorders eventually result in walking impairments [1], [7]. Due to unbalanced walking, MS patients have to change their walking posture to maintain walking stability. More than 90% of MS patients have lower limb movement problems [42]. It has been reported that patients have reduced speed and cadence, step length and stride length, and increased base width, step time, stride time, and double support time compared to healthy individuals [43].

2.2.2 The Assessment of MS

As the reaction of patients to rehabilitation training is unique for each individual, designing effective evaluation methods and walking tests to measure MS patients' walking ability before and after rehabilitation is critical in medical examinations. Expanded Disability Status Scale (EDSS), Multiple Sclerosis Impact Scale (MSIS-29), Self-Selected speed walking test (SS), Dual-Task walking test (DT), and bipedal hopping tests are possible ways to measure and analyze walking patterns in patients with MS.

Expanded Disability Status Scale (EDSS), a widely used assessment tool for MS patients, is a method to describe the disability level of MS to assess the patient's disability changes over time [8]. Neurologists usually measure EDSS by looking at muscle weakness, ability to move arms and legs, uncontrollable eye movements, unusual sensations or numbness, and other functions. The EDSS score ranges from 0 to 10. Level 4 of EDSS is the starting point of significant walking impairment, and level 6 indicates that patients need assistive devices to help

them walk [7]. Although examining the EDSS score for each patient is a standard process when profiling this disease, EDSS has been reported to have limitations. For example, the EDSS score might be intensely subjective due to the assessment process. In addition, EDSS scoring is not always evenly distributed. After level 4, EDSS seems to focus more on walking disability than on other aspects of the disease [44]. Therefore, although EDSS is widely used in MS assessment, its use in patients with a mild disability is not sufficient to accurately describe the disability. However, machine learning can evaluate the difference between the disability levels of patients by analyzing the relationship between gait measurements and the patient's walking ability.

Multiple Sclerosis Impact Scale (MSIS-29) is another scale used by clinicians to assess the disease severity by asking patients to self-report several questions related to physiological and psychological functions. Unlike EDSS, the MSIS-29 focuses on the patient’s perception of the impact of MS on their walking and cognitive abilities [10]. MSIS-29 contains nine cognitive-related and 20 physiological-related questions. Each question is scored from 1 (not affected) to 5 (severely affected) [9]. Thus, MSIS-29 presents a patient’s perceptions of the impact of the disease on cognitive and physiological aspects before neurologists detect the changes clinically.

Self-Selected speed walking test (SS) is a widely used clinical walking test that simulates daily walking. The SS requires participants to walk at their comfortable pace, with researchers recording the gait measurements simultaneously. While walking at a comfortable speed, the energy cost can be used to differentiate healthy people from patients with MS. Stella et al., in 2020, examined the energy cost differences between healthy controls and MS patients. The result showed that the energy cost of MS patients was 2 to 3 times higher than that of healthy

controls while walking at a comfortable speed [45]. SS is now prevalently used to detect the possibility of falls or to predict daily activities among older adults [46], [47]. SS was also used to detect the relationship between gait variables and the EDSS score. Preiningerova et al., 2015, asked their MS patients to perform the SS test to assess the changes in walking ability that align with the EDSS level. In their study, patient's EDSS scores ranged from 0 to 6.5, where patients can walk with or without walking aids. Totally 284 MS patients participated in the study. The researchers found that, as the EDSS score increased, the patients had increased step length, prolonged step time, lower velocity, and extended double support time [48].

The Dual-Task walking test (DT) requires patients to walk while performing simple cognition tasks simultaneously, such as subtracting 7 starting from 100 or walking while talking. By applying cognitive tasks while walking, gait measurement differences between healthy people and patients may reflect cognitive impairments in patients. To apply DT in the long run, its reliability must be tested. The validity of the walking-while-talking test in detecting minimal detectable changes has recently been attested by Henning et al. [49]. In 2020, Chen et al. examined the reliability of various DT parameters over the long term. Eighteen MS patients and 12 healthy controls were involved in the study. This study was designed to examine the reliability of SS, DT variables, and cognition over three sessions. Dual-task cost, the percentage changes in ST and DT, were also calculated for each variable. The greater the dual-task cost, the worse the performance was. The gait variables were obtained using a walkway. The authors concluded that the dual-task cost and coefficient of variability were unreliable; however, the pace and cognition rates showed outstanding reliability [21]. Kirkland et al. (2015) examined

different types of DTs to determine the most reliable DT test for disease steps assessment, which is a rating scale for functional disability level [50]. Patients performed three DT tasks: reciting every second letter of the alphabet, counting while leaving out the number including three, and repeatedly subtracting seven from 100. The third method was the best as it presented the most consistent cognitive dual-task cost for stride length and disease steps [51].

Bipedal hopping is an ambulatory test that can detect the muscle strength of a patient. It is a more challenging test to detect patients' subtle deficits. Patients who pass the previous walking test may fail this test [52]. Kirkland et al. timed their participants bipedal hopping to a metronome to detect anticipatory motor control impairments. This study included 13 MS patients, 9 matched healthy controls, and 13 elderly controls. Participants performed 40 beats/min and 60 beats/min of bipedal hopping following the metronome. Hop-related parameters (length, symmetry, and variability) were extracted to test the motor control ability. The results indicated that MS patients took a longer time to react to the metronome than the elderly and healthy controls. People with MS and healthy controls had a longer hop length, wider hop width, and higher velocity than the elderly group at 60 beats/min. The difference between MS patients and the elderly was much smaller at 40 beats/min [38].

Kirkland et al. used bipedal hopping to reveal evidence of advanced neuromuscular aging in patients with mild multiple sclerosis. This study used bipedal hopping to examine whether patients with mild MS had signs of neuromuscular aging. MS patients (n = 13), matched controls (n = 9), and the elderly group (n = 13) participated in this study. Participants first performed a timed twenty-five-foot walk test. After rest, they underwent a bipedal hopping test.

The researcher extracted lower limb power, consistency, and symmetry parameters as features to draw conclusions. They found that MS patients had the highest integral pressure, longest hop time, and percentage stance time, and slowest velocity compared with the elderly and controls. For hopping consistency, the record for the variability of each characteristic patients fell between those of the controls and the elderly group [39]. These studies proved that bipedal hopping was able to detect athletic ability and motor prediction impairment in patients with MS. However, such tests require the hopping ability in patients, making it difficult to examine subtle changes in patients with higher disability levels.

2.2.3 Gait Analysis

Gait analysis is now prevalent in the medical field and is a valuable method to help clinicians and therapists analyze patients' current gait disorders and walking disabilities over time. Researchers from different backgrounds have conducted gait analyses on gait measurements using various techniques.

Statistical analysis is a widely used method for examining the differences and relationships between different groups. Preiningerova et al. conducted statistical analyses to analyze frequently used spatial and temporal gait measurements of MS patients, with the goal of determining which parameters are the best in determining the different disability levels of patients. In this study, all variables of each group are presented as mean ± standard deviation, and the differences between each group are compared [48]. Kim et al. applied the ANOVA test with Bonferroni adjustment to analyze the kinematic and kinetic features to determine the

differences in walking between the elderly and young people [53]. Although statistical analysis is helpful in determining the different patterns of various groups, it is possible that the variations between individuals decrease when all values are combined to calculate them for a summary result. Newer technologies apply machine learning to gait measurements to describe the relationship between gait variables and disease.

2.3 Machine Learning

Machine learning involves learning from data. The main idea of machine learning is to discover unobservable patterns by analyzing existing data using suitable algorithms [54]. In practice, machine learning can be divided into six steps: collecting data, pre-processing the data, choosing a model, training the model with the data, assessing the result, and using the model for prediction [54]. Machine learning algorithms can be broadly categorized into supervised and unsupervised learning.

2.3.1 Supervised Learning

Supervised learning uses labeled data to train machine learning algorithms to predict or classify data. The labeled datasets can be separated into training and testing sets. The supervised learning models are fitted to the training set to learn the patterns between the input and output in the training data. The trained model is then applied to the testing set to assess the model performance by comparing the prediction result with the data label. Supervised learning can be divided into classification and regression. Classification aims to assign data to specific classes,

and the goal of regression is to calculate the relationship between the variables and the prediction target [55].

K-nearest neighbors (kNN) is associated with simple instance-based models. This machine learning model learns only from existing samples. After being trained with the existing variables and classes, the trained model assigns the sample to the closest class when it encounters new samples. Here, k refers to the k-nearest neighbors searched when the algorithm comes across a new sample and assigns the new sample to the group with the majority number of neighbors. To determine the closest group, the distance between the new sample and the existing data must be calculated. The Euclidean distance between data points is calculated using Eq 1 with x and x' representing two points in Euclidean n-space; x_i and x_i' represent Euclidean vectors. The kNN is easy to implement and requires only tuning of the k value.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2} \quad (1)$$

Logistic regression (LR) [56], despite the term regression, is arguably a popular classifier in machine learning used for binary, one vs. rest, or multinomial classification. It provides a probability for the target class between 0 and 1 to describe the relationship between the input variables and one or more output targets. The algorithm relies on a logistic function into which input values x are fitted, where the weights or coefficient values are adjusted to predict the output value y. The standard logistic function is represented by Eq 2.

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

Support Vector Machine (SVM) effectively works with high-dimensional data [57]. SVM defines a hyperplane boundary in an N-dimensional space, where N is the number of input

features. As many hyperplanes exist in this space, SVM attempts to find the optimal plane that maximizes the separation of both classes. Additional points can then be classified as belonging to either class depending on the side of the optimal hyperplane they occupy.

Extreme Gradient Boosting (XGB) [58] is an optimized distributed gradient boosting library introduced by Chen and Guestrin in 2016. The gradient boosting technique is used for both classification and regression. Applied to an ensemble of weak prediction models, such as decision trees, the gradient boosting algorithm trains data with these weak learners, forces the poor learners to learn to increase their prediction score, and finally combines them into one accurate prediction algorithm. XGB utilizes the concept of gradient tree boosting while introducing regularization parameters to reduce overfitting.

AdaBoost, which is short for adaptive boosting, is a machine learning method that focuses on previously under-fitted samples. AdaBoost first trains a base classifier with the original dataset and then uses the base classifier model to predict the training set. When the prediction is completed, the model training focuses on the samples that were not accurately predicted by the base classifier, and this process is repeated until the final model best fits the dataset [59].

2.3.2 Unsupervised Learning

Unsupervised learning is another machine learning technique that uses unlabeled data [60]. This learning method clusters information in the dataset and finds unique patterns to separate the data into distinct groups without labeling the data. As unsupervised learning discovers information in the input data, it is helpful to find unknown patterns in input variables to

customize segmentation, for example, footprint clustering from pressure sensors, as demonstrated in this thesis.

2.3.3 Machine Learning used for MS Gait Analysis

Machine learning is a useful tool for gait analysis, as it can deal with an enormous number of gait variables, discover unobserved patterns between gait measurements and the disease, and apply these patterns to predict the future. McGinnis et al. (2017) designed a study to estimate the gait speed by applying machine learning to skin-mounted wearable sensors. The sensors were attached to the skin of the participants' sacrum, both thighs, and both shanks. Thousands of data points were generated from the sensors, and support vector regression models were used to estimate walking speed according to the data points. Using machine learning, gait speed was accurately estimated by analyzing the features of skin-mounted wearable sensors [61].

Gait studies using the machine learning method were also conducted by researchers. In 2018, Supratak et al. conducted a study to validate whether a timed 25-foot walk test could predict the real-life walking speed. Thirty-two patients with EDSS scores ranging from 0 to 6 were recruited. They were required to wear an accelerometer for up to 7 days. In-home walking gait speed was calculated. Principal component analysis was conducted on the features extracted from the accelerometer to compute the first and second most important components. The results of the trained model demonstrated that the walking speed from the timed 25-foot walk test matched the maximum sustained walking speed in-home environment [62].

Machine learning has also been used to analyze fatigue in patients. Ibrahim et al. in 2020 used machine learning algorithms to reveal the relationship between gait parameters and patient-reported fatigue. An inertial measurement unit was used for each foot of each participant. Researchers used the Borg scale of self-perceived exertion for fatigue measurement during exercise. The scale ranged from 6 (no exertion) to 20 (maximum exertion) [63]. Principal component analysis was first performed to obtain the principal components within the dataset, and a random forest regressor was then applied to these components to estimate the fatigue value. The results indicated that machine learning could predict fatigue by analyzing the temporospatial features [64].

Besides predicting gait speed or self-reported fatigue from patients' data, Kaur et al. in 2020 explored whether machine learning algorithms can help distinguish older patients with MS from healthy controls. Gait measurements were acquired from an instrumented treadmill and analyzed using nine machine learning models, decision tree, random forest, support vector machine with linear and radial basis function kernels, gradient boosting machine, adaptive boosting, extreme gradient boosting, multilayer perceptron, and logistic regression. Gradient boosting and multilayer perceptron were identified that surpassed the others. The results also proved that machine learning can distinguish aging patients from healthy controls [65].

2.3.4 Model Evaluation

A cross-validation technique was used to test model performance. The cross-validation requires the original dataset to be divided into training and testing sets. As the name implies,

the algorithm uses the training set data to train the models and the trained model to predict on the testing set. The model performance can be calculated using the prediction results and testing data labels [30]. Precisely, k-fold cross-validation randomly separates a dataset into k-folds that contain an equal number of samples. One fold is set as the testing data, and the remaining folds serve as the training data. The model was trained with the training set, tested using the testing data, then another fold was set as the testing data, and the process was repeated until all the data samples were tested.

Ibrahim et al. used nested 10-fold cross-validation as a model performance assessment method [64]. The dataset used in this study includes 32 women and 17 men whose EDSS scores range from 1 to 6.5. Some of the gait features used in this study and this thesis overlap, which are stride length, gait velocity, stride time and stride length. The nested 10-fold cross-validation sets the k-fold cross-validation for hyperparameter selection as part of the k-fold cross-validation of the model selection. Therefore, it is also referred to as double cross-validation. This cross-validation method reduces the bias during hyperparameter optimization and the likelihood of model overfitting.

Leave one subject out cross-validation is another kind of k-fold cross-validation with $k=n$, where the n is the total number of samples. Leave one out method uses the data of all but one participant's data as the training set while setting that one patient's data as the testing set. The process was repeated until all participants were used as testing data. By testing the unobserved data, this process ensures that the model provides accurate predictions [30].

3 Methodology

This chapter describes data preprocessing methods. The design of MS vs. Healthy control classification and mild MS vs. moderate MS classification, including machine learning algorithms and cross-validation method were also introduced. Methods for feature selection and feature importance calculation were also presented. Python 3.7 was used to implement the feature creation and machine learning process.

3.1 Data Collection and Experiments

3.1.1 Data Collection

Data were collected as part of the Health Innovation Team in MS (HITMS) project, which is a longitudinal study of people's health with MS in Newfoundland and Labrador, Canada [66], [67]. The study was approved by the Institutional Health Research Ethics Board (HREB # 2015.103). Raw walkway sensor data were collected from participants in this study between 2016 and 2020. Each patient had at least one visit and was permitted to walk with or without a walking assistive device [68]. Controls also had at least one visit and were required to have no gait disorders.

Demographic data, including age, height, and weight, were gathered for all patients. People with MS were diagnosed by an MS neurologist, and their disease severity was scored using EDSS [69]. The EDSS score ranges from 0 to 10; a score of 0 indicates no observable gait dysfunction in patients; a score of 6 indicates that the patient requires bilateral walking aids and can walk at least 20 m; and a score of 10 indicates death due to MS. Patients in this project had

EDSS scores no higher than 6.5 because patients with EDSS scores higher than 6.5 were not able to finish walking tests across the walkway. Thus, data from 107 patients and 16 healthy controls data were included in the study. The average EDSS score of 107 patients was 2.11 ± 1.89.

All the patients were required to complete the MSIS-29 questionnaire. The MSIS-29 is a standardized self-evaluation form that asks patients to rank the impact of MS symptoms using various physical and psychological questions [70]. The MSIS-29 score ranges from 1 to 5, with 1 indicating no impact, 2 indicating a little impact, 3 indicating moderate impact, 4 indicating quite a bit impact and 5 indicating severe effects on a patient's lifestyle [71]. The replies to a subset of MSIS questions on how patients felt about their movement, as indicated in Table 3.1, were used to determine the patient exclusion criteria.

An instrumented walkway (Zeno Walkway, Protokinetics Haverton PA) was introduced to this project. The walkway measured 90 cm x 420 cm and was embedded with a matrix of pressure sensors. Spatial measurements were provided as the (x, y) positions of the activated sensors and converted to distances measured in centimeters. Each sensor has an area of 1.27cm x 1.27 cm [72]. Timestamps were recorded when sensors detected ambulation, and the corresponding pressure levels were recorded in milliseconds.

All the participants were required to complete two clinical walking tests across the walkway. The first walking test was the Self-Selected speed walking test (SS). All the participants walked at a comfortable pace. The second was the Dual-Task walking test (DT), with the participants required to walk along the walkway while subtracting 7 from 100. Each

walking test required the participants to walk at least two passes on the walkway; therefore, at least ten steps were recorded.

3.1.2 Experiment 1. Classifying MS Patients versus Healthy Controls

In this experiment, binary classifiers were trained using data from healthy controls and patients with at least one MSIS-29 score equal to or greater than 3. This experiment aimed to determine whether machine learning models could distinguish patients from healthy controls using a standard feature set and whether the newly designed features calculated from the raw sensor data (see Section 3.2.3) could further improve model performance.

Patients with a score of 3 or higher for at least one of the MSIS-29 questions were selected for this experiment. Thirty-five patients were excluded from this study. The average EDSS score of the remaining patients was 2.73 ± 2.04. Healthy controls were not required to complete the MSIS-29 questionnaire or EDSS examinations. The final dataset for this experiment included 72 patients and 16 healthy controls (Table 3.1).

Table 3.1 Patient demographic data and MSIS-29 score of MS vs. healthy control classification.

Patient Data Features	**Mean, Variance**
EDSS Score	2.73 ± 2.04
Age (years)	48.40 ± 9.95
Height (cm)	169.82 ± 8.64
Weight (kg)	83.52 ± 23.38
MSIS-29-Q4 Problems with your balance?	2.99 ± 0.94
MSIS-29-Q5 Difficulties moving about indoors?	2.14 ± 1.01
MSIS-29-Q6 Being clumsy?	2.76 ± 1.01

MSIS-29-Q7 Stiffness?	2.86 ± 1.15
MSIS-29-Q8 Heavy arms and/or legs?	2.90 ± 1.14
MSIS-29-Q9 Tremor of your arms or legs?	2.17 ± 1.17
MSIS-29-Q10 Spasms in your limbs?	2.29 ± 1.25
MSIS-29-Q11 Your body not doing what you want it to do?	2.39 ± 1.21

3.1.3 Experiment 2. Classifying Mild Patients versus Moderate Patients

This experiment aimed to determine whether machine learning models could detect the level of self-reported gait disorders which were used to distinguish mild patients from moderate patients. Two sub-experiments were conducted to determine the machine learning baseline score of the standard feature and whether the augmented feature set could improve model performance.

No data from healthy controls were included in the mild/moderate patient classification. The MSIS-29 was used to determine the patients' gait disorder levels as the gold standard for classification. Patients who scored less than 3 while answering the MSIS-29 were considered mild. The remaining patients were considered moderate. Data from all 107 patients were analyzed (Table 3.2).

Table 3.2 Patient demographic data and MSIS-29 score of mild MS vs. moderate MS patient classification.

Patient Data Features	**Mean, Variance**
EDSS Score	2.11 ± 1.89
Age (years)	47.58 ± 9.98
Height (cm)	168.05 ± 8.06
Weight (kg)	81.61 ± 20.37

MSIS-29-Q4 Problems with your balance?	2.47 ± 1.08
MSIS-29-Q5 Difficulties moving about indoors?	1.80 ± 0.96
MSIS-29-Q6 Being clumsy?	2.30 ± 1.08
MSIS-29-Q7 Stiffness?	2.38 ± 1.18
MSIS-29-Q8 Heavy arms and/or legs?	2.35 ± 1.23
MSIS-29-Q9 Tremor of your arms or legs?	1.81 ± 1.08
MSIS-29-Q10 Spasms in your limbs?	1.90 ± 1.16
MSIS-29-Q11 Your body not doing what you want it to do?	2.00 ± 1.14

3.2 Data Analysis and Feature Extraction

3.2.1 Deriving Footprints from Raw Sensor Data

The attributes of the raw walkway sensor data included time, x-coordinate, y-coordinate, pressure level, foot type, foot count, footfall, and pass index. For each step, the maximum pressure reading for each sensor was involved in building footprint. After footprints were built, spatial centroid of the footprint was set. The x and y coordinates and timestamp data allowed the reconstruction of each pass across the walkway.

Raw data were partitioned into individual footfalls using a k-means clustering [73] for each pass recording. This unsupervised clustering algorithm separates the *n* spatial coordinates into *k* individual footfalls, where each observation belongs to the cluster with the nearest centroid. The *k* was empirically determined as 5 for each pass. When doing walking tests, the walkway system records gait data and provides instant replay. If a patient walked out of the walkway accidentally or patients walked with a shuffling gait, such non-standard walking steps were excluded using the walkway system. Walking data were then clustered using k-means. The

footprint images were manually checked after the clustering to make sure the final dataset included only correct steps.

For each footprint data (matrix), a quadrilateral was used to enclose the foot shape. The quadrilateral was generated using four lines: a medial line connected to two outer sensors on the medial side of the footprint, a lateral line connected to two outer sensors on the lateral side of the footprint, a rear line perpendicular to the medial line starting from the rear outer sensor, and a front line perpendicular to the medial line starting from the front outer sensor [72]. This quadrilateral was then subdivided into three regions with individual sub-centroids, which provided further details on the heel, mid, and fore of the footprint. Figure 3.1 demonstrates how a footprint is segmented.

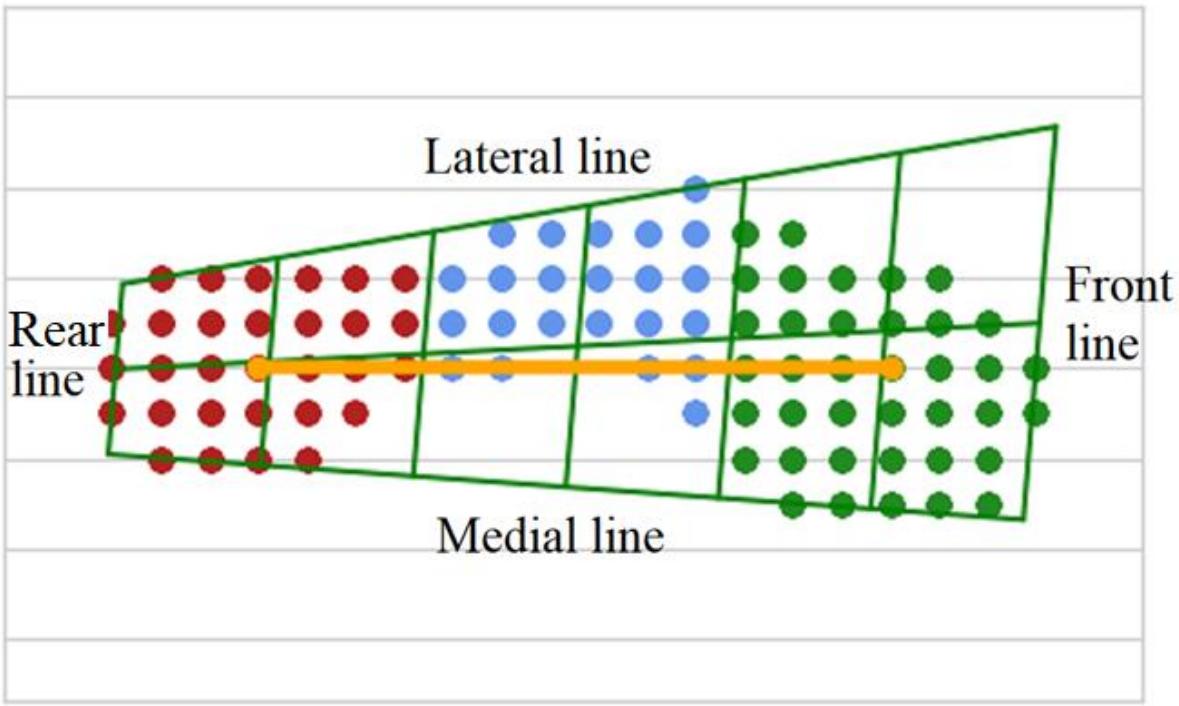

Figure 3.1 Footprint showing heel (red), mid (blue), and fore (green) sections, as well as the centerline of the foot (yellow). The segmented quadrilateral encloses the shape.

3.2.2 Standard Gait Features

After identifying unique footfalls in the gait recording, a set of standard gait features was extracted based on each pass. These included the step/stride length and width, toe angle, step/stride time and velocity, single/double support time, and stance time.

The details regarding each parameter can be found in Table 3.3.

Table 3.3 Detailed description for each standard gait parameter

Standard Gait Features		
Spatial Features	Foot Type	A descriptor for right or left foot.
	Foot Length (cm)	Measured as the distance between heel/fore centroids of the same foot multiplied by 1.5.
	Foot Width (cm)	Measured as the distance across the midpoint of the subregion enclosing the fore section of the footprint.
	Foot Area (cm^2)	Measured as the total activated area of the sensors involved in generating the footprint.
	Toe Angle	Measured as the angle between the line of progression (LOP) (the line connecting the heel centers of two consecutive footprints of the same foot) and the midline of the footprint (the line connecting the heel and fore centroids of a given foot).
	Step Length (cm)	Measured along the direction of the walkway, from the heel center of the current footprint to the heel center of the previous footprint on the opposite foot.
	Step Width (cm)	Measured from the midline midpoint of the current footprint to the midline midpoint of the previous footprint on the opposite foot.
	Stride Length (cm)	Measured on the LOP between the heel points of two consecutive footprints of the same foot.
	Stride Width (cm)	Measured as the vertical distance from the midline midpoint of one footprint to the line formed by the midline midpoints of two footprints of the opposite foot.
	Base Width (cm)	Measured as the vertical distance from the heel center of one footprint to the LOP formed by two footprints of the opposite foot.
Temporal Features	Step Time (sec)	The time elapsed from the first contact of one foot to the first contact of the opposite foot.

	Stride Time (sec)	The time elapsed between the first contacts of two consecutive footfalls of the same foot, measured in seconds.
	Stride Velocity (cm/sec)	Obtained by dividing the stride length by the stride time.
	Step Velocity (cm/sec)	Obtained by dividing the step length by the step time.
	Single Support Time (sec)	The time between the last contact of the current footfall to the first contact of the next footfall of the same foot.
	Double Support Time (sec)	The time between the heel contact of the next footfall to toe-off of the current (and opposite) footfall.
	Stance Time (sec)	The time between first contact and last contact of the same foot.

3.2.3 New Feature Design

New parameters, presented in Figure 3.2 and explained in detail in Table 3.4, were also designed and calculated from the raw sensor data. To the best of our knowledge, these features have not yet been rigorously tested in machine learning classification settings.

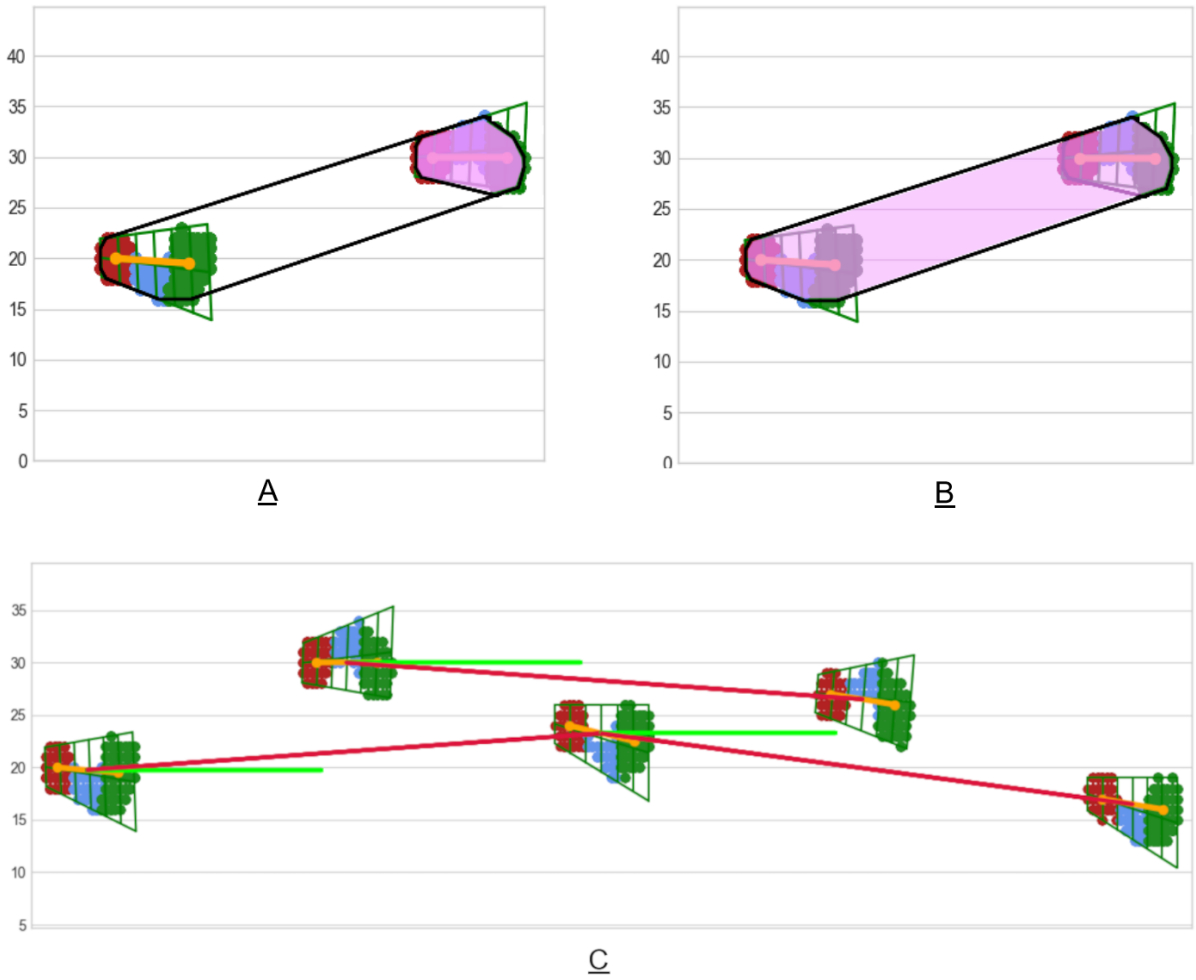

Figure 3.2 A. The pink shaded region shows the base of support area between two successive footfalls. B. The pink region shows the hull area for a single footfall. C. The green line represents the expected LOP, and the red line represents the actual LOP between two consecutive footfalls of the same foot. The angle between the desired and actual lines is the line of progression deviation angle.

Table 3.4 Detailed descriptions of newly designed features

New Feature Design	
Toe Direction	Standard toe angle [74] on the walkway is recorded as a signed value. The original toe angle value is split into two features: magnitude and direction. The direction of the angle indicated that whether the foot is toe in or toe out. The magnitude is the value of the angle.
Hull Area	To better approximate the actual shape of the footprint, the convex hull enclosing the point cloud was calculated for each footprint. The hull area is the area enclosed by the line segments bounding the footprint tightly in a convex hull. Figure 3.2A shows the hull area.

Base of Support (BOS) Area	In gait, the BOS [75] is commonly measured as a one-dimensional length. A convex polygon constructed to enclose two footprints (a footprint and its preceding print) used to approximate the area of the BOS. Figure 3.2B shows the picture of the BOS area.
LOP Deviation Angle	The deviation angle between an expected regular LOP and the actual LOP. The actual LOP is the line connecting the heel centers of two consecutive footfalls of the same foot. Ideally, the patient should walk parallel along the walkway. Figure 3.2C shows the picture of the LOP deviation angle

3.2.4 Machine Learning Feature Sets

Two feature sets, the standard and augmented sets, were designed for the experiments. The purpose of conducting experiments with these two feature sets was to assess the potential of the newly designed or unutilized features to determine the utility of creating new features from raw walkway data.

The standard set included 14 features: step time, stride time, step velocity, stride velocity, single support time, double support time, stance time, foot type, signed toe angle, step length, step width, stride length, stride width, and base width.

Foot length, width, and area are rarely documented as useful features in gait-related classification studies. Therefore, in this thesis, these features were included and used in the augmented feature set (see Section 3.2.4) to examine whether they can further improve classification accuracy. The augmented set included all features from the standard set and seven additional parameters, foot length, foot width, foot area, hull area, LOP deviation angle, unsigned toe angle, and toe direction.

3.3 Machine Learning Process

3.3.1 Data Cleaning and Missing Data Prediction

Data cleaning is the process of removing or fixing missing, irrelevant, duplicated, or incorrect data from a dataset. The goal was to improve data quality because using data that included misinformation would lead to erroneous results in interpreting the answer to research questions [76].

Missing data are another problem when cleaning data, and two methods can be used to address this issue. The first is to remove the rows with missing data when the size of the dataset, after removal, is sufficiently large for training [77]. In this project, the first and last steps did not have the step/stride length and width data or LOP angle because of the nature of the walking tests. Samples with missing data were excluded from the analysis. Another method is to replace the missing data with reasonably predicted values using machine learning strategies if the missing data points are removed, the dataset is not large enough for training [16]. Data points in the dataset with complete information were used for model training. The data points with missing data were used as the prediction dataset. The machine learning model was first applied to data points with complete information. The trained model was then applied to the prediction dataset, and the missing data were replaced with the predicted result. The participant's height for data normalization was calculated using this method. The process is described in detail in Section 3.3.3.

The features were directly collected and calculated from walkway sensor data. Misreported sensor points or footsteps that were too close to each other, leading to the miscalculations by of k-means cluster, were excluded. Data outliers, such as the negative value of time-related features or the value of spatial features that were two or more times higher than the average value, were deleted. The original dataset contains 5931 samples, and samples refer to footprints. For each participant, 45 to 50 samples were obtained. After data cleaning, the final dataset of MS vs. Healthy control classification has 2217 samples, and the final dataset of mild MS vs. moderate MS classification has 2317 samples.

3.3.2 Data Balancing

Imbalanced data are a common problem when in machine learning classification, meaning that the number of samples belonging to each class in the problem is not evenly distributed, with some classes having a much higher number of samples than others. This situation could lead to the bias of the prediction result. For example, if 90% of the data belongs to the same class, the model will reach 90% accuracy by classifying all data into the same class, but the model is biased. Therefore, data balancing is required to obtain an unbiased model. In our case, the patient/control ratio was about 6:1, and the mild/moderate patients had a ratio of 1:5. Therefore, during data analysis, data balancing was performed for both experiments prior to proceeding with the classification analysis [77]. The synthetic minority oversampling technique (SMOTE) [78] and adaptive synthetic sampling approach (ADASYN) [79] were introduced as data balancing methods in this project, and the one that provided the best model performance

was selected for the further classification. SMOTE synthesizes a new sample by randomly choosing a data point from a line segment in the feature space, formed by a minority class sample m and one of the m's k-nearest neighbors (usually k = 5, both randomly chosen). This process is repeated until the data of the two classes are balanced. The ADASYN was first introduced by He et al. (2008). This data balancing method works in a manner similar to SMOTE. However, the new samples generated by this method aim to create synthetic data next to the minority class samples, which are incorrectly classified by the k-nearest neighbors [79]. The bias is thus reduced, and such a method helps the model to learn difficult minority samples. MS vs. Healthy control classification worked better with the SMOTE method, whereas better results were obtained for mild MS vs. moderate MS classification with ADASYN.

3.3.3 Data Normalization

Data normalization involves rescaling the data in different numeric columns/features to a similar range of values without distorting the distribution in each column. Machine learning estimators may perform poorly if the features are not normalized. The numerical data collected in this thesis exhibited a variety of ranges and thus required rescaling. After normalization, the features were scaled to exhibit zero mean and unit variance, except for foot length, width, area, and hull area. In general, taller people have larger foot length and wider foot width. The foot length, width, area and hull area were then scaled according to the height of the participant [80]. This project accomplished the normalization process of foot length, width, area, and hull area by dividing each participant's parameter measurements by the participant's height (cm). For

participants with no record of height data, SVM was used to predict the value of the missing data. Participants with complete data were used to train the SVM regression model. The height of these participants was set as a prediction label, and their foot length, foot width, foot area and hull area as features. The trained SVM model was then applied to predict the height of patients who had no record of height. The foot length, foot width, foot area and hull area of these participants were used as the input data for their height prediction.

3.3.4 Feature Selection

The original dataset includes various features. Some were useful for classification, while others might provide little or no helpful information to distinguish between classes. It is necessary to exclude these less-helpful features to reduce overfitting, increase training speed, limit unnecessary noise in the data, and improve model performance.

A straightforward way to select features is to calculate the correlation between the features as well as between the features and the target. Highly correlated features provide similar information to the model and thus can be reduced using feature selection or other techniques. Pearson's correlation was calculated for both experiments to reduce the number of dependent features, and a correlation matrix with a heatmap [81] was used to visualize the correlations between features. The Seaborn library was used to plot a heatmap. Seaborn is a visualization library that provides an interface for drawing statistical graphics [82]. The relationship between the two features can be positive, negative, or none, as indicated by the color of the heatmap. The resulting feature correlation matrix consists of scores ranging from -1, strong negative

correlation, to +1, strong positive correlation, with a score of 0 indicating no correlation between the features. Our study used a removal threshold of -0.8 / 0.8 for feature selection, whereby features scores higher than 0.8 or lower than -0.8, would be considered for removal.

The remaining features were tested using Analysis of Variance F-Test statistics (ANOVA) [83] and recursive feature elimination with cross-validation (RFECV) separately to form a subset of numerical features that had the greatest impact on the model result. The method that provides better model performance was selected. The ANOVA assigned each feature a score. The higher the score, the stronger is the features with more significant unexplained variance in the prediction. When the features were ranked by their F-statistic scores, the size of the optimal feature set had to be determined. For each possible size s_i of the final set [1, 2, ...n features], a grouped 5-fold cross-validation strategy [84] with an SVM classifier was used to obtain prediction accuracy of each size s_i. The average prediction accuracy was collected for each size s_i, and the optimal size was chosen based on the highest score.

Because categorical features were not included in the correlation or feature selection process, they were reintroduced into the final feature set after the numerical feature selection process was completed.

RFECV selects the final feature set by recursively eliminating features from the dataset according to the cross-validation score of an external estimator and the feature score. RFECV uses two outputs to measure the feature score: feature correlation and feature importance of the estimator. Cross-validation within the RFECV is used to calculate the model performance. Features with lower scores were excluded. The selection process of RFECV stops when the

final set of features helps the model obtain the best performance matrix. In this study, XGB was used as an external estimator. Both experiments performed better with the ANOVA-SVM feature selection method.

3.3.5 Feature Importance

Feature importance can be calculated for the LR and XGB classifiers. The ANOVA-SVM feature selection process also calculates the average F-score for the selected features. Linear models, such as logistic regression or support vector machines with linear kernels, calculate the coefficient of each feature to determine the feature importance. Tree-based models, such as random forest or XGB, calculate feature importance to explain the model results. Feature importance was calculated based on how the nodes of the tree used in the training improved the model results. The scores of the features were automatically computed when training was completed. All the scores were scaled, and their sum was equal to one. The higher the score, the higher is the importance of the feature. The ANOVA-SVM process calculated the average F-score for each selected feature. In this project, the average SHapley Additive exPlanations (SHAP) value [76] of each feature was calculated to determine the importance of each feature to the model.

3.3.6 Machine Learning Algorithms

Logistic regression (LR) [56], support vector machine (SVM) [57], and extreme gradient boosting (XGB) [58] were selected as they represent three well-known classification methods: probability, hyperplane polarity, and boosted decision-tree ensembles, respectively. Given a set

of features, each model was studied to categorize the footprint as belonging to an MS patient or a healthy control (Experiment 1) and to distinguish mild MS patients from moderate MS patients (Experiment 2) using a range of classification scoring metrics.

3.3.7 Training and Evaluation

To further reduce overfitting, a grouped 5-fold cross-validation strategy was employed. All data points in the training set were grouped according to the date of the participant visit and their unique ID. Each group was assigned a unique identifier for the separation. These groups were randomly and evenly split into five folds. For each round of validation, four folds were set as the training data set, and the remaining fold was set as the testing data. The models were trained on the training data and evaluated using testing data. This process was repeated until all the five folds were tested. The groups remained intact throughout the training/test validation splitting, and no group was permitted to appear in two different folds. In this fashion, the data of the same participant were not used simultaneously in the training and testing sets to avoid data leakage between these two datasets.

Each model in this study had a unique set of hyperparameters that required tuning to provide the best results. A standard grid search method was used to test each model across a range of hyperparameter settings. The setting that provided the best scores were then selected. A summary of the tested parameters for each model, along with the optimal hyperparameter settings selected by the grid search, is presented in Tables 3.5, 3.6 and 3.7.

Table 3.5 Hyperparameter options for each model

Algorithms	Hyperparameter Options
LR	'solver': ['newton-cg','lbfgs', 'liblinear'], 'penalty': ['l1', 'l2', 'elasticnet'], 'C': [100, 10, 1.0, 0.1, 0.01], 'max_iter': [200, 400, 600]
SVM	'kernel': ['poly', 'rbf', 'sigmoid', 'linear'], 'C': [8, 7, 6, 5, 4, 3, 1.0, 0.5, 0.1], 'degree': [0, 1, 2, 3, 4, 5]
XGB	'max_depth': [2, 3, 4, 5, 6], 'eta': [0.1, 0.2, 0.3, 0.4], 'objective': ['binary:logistic', 'binary:logitraw', 'binary:hinge']

Table 3.6 Optimal hyperparameters for MS vs. healthy control classification.

Algorithms	Optimal Hyperparameters for Standard Set	Optimal Hyperparameters for Augmented Set
LR	'C': 1.0, 'penalty': 'l2', 'solver': 'newton-cg'	'C': 1.0, 'penalty': 'l1', 'solver': 'liblinear'
SVM	'C': 3, 'degree': 3, 'kernel': 'rbf'	'C': 5, 'degree': 3, 'kernel': 'rbf'
XGB	'eta': 0.3, 'max_depth': 3, 'objective': 'binary:logistic'	'eta': 0.3, 'max_depth': 3, 'objective': 'binary:logitraw'

Table 3.7 Optimal hyperparameters for mild MS vs. moderate MS patient classification.

Algorithms	Optimal Hyperparameters for Standard Set	Optimal Hyperparameters for Augmented Set
LR	'C': 10, 'max_iter': 200, 'penalty': 'l2', 'solver': 'liblinear	'C': 10, 'max_iter': 200, 'penalty': 'l2', 'solver': 'liblinear'
SVM	'C': 6, 'degree': 0, 'gamma': 'scale',	'C': 8, 'degree': 0, 'gamma': 'scale',

	'kernel': 'rbf'	'kernel': 'rbf'
XGB	'eta': 0.1, 'max_depth': 4, 'objective': 'binary:logistic'	'eta': 0.05, 'max_depth': 6, 'objective': 'binary:hinge'

The number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [85] predictions were calculated for each model, and the accuracy, precision, recall and F1 score were calculated to gauge the model effectiveness (Table 3.8).

The terms positive and negative represent two different targets. For example, patients (moderate patients) were considered positive. Healthy controls (mild patients) were considered negative. TP means that the model predicts the data as 'patient' and the data really belong to a 'patient'. FP means that the model predicts the data as 'patient', but the data actually are 'control'. The same applies to the negative sets.

Table 3.8 Details for accuracy, precision, recall and F1-score.

Accuracy (%):	The total number of correct predictions out of the total number of all predictions.	$\frac{TN+TP}{TN+TP+FN+FP} * 100\%$
Precision (%)	Positive Predictive Value. Represents the proportion of true positive predictions to all actual positives.	$\frac{TP}{TP+FP} * 100\%$
Recall (%)	Sensitivity. Measures the proportion of true positives to the total number of actual positives.	$\frac{TP}{TP+FN} * 100\%$
F1 Score (%)	The weighted harmonic mean of precision and recall.	$2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} * 100\%$

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were also generated for each model. The area under these curves (AUC) can be assessed as another

measure of determining the model's predictive capability. The larger the area, the better the model performance is.

4 Experimental Results

This chapter presents statistical analysis for participants and results of MS patients vs. healthy control classification and mild MS vs. moderate MS classification, respectively.

4.1 Statistical Analysis

Table 4.1 Mean values of gait parameters for mild patients, moderate patients, and healthy controls.

Parameters	**Mild Patients**	**Moderate Patients**	**Controls**
Number	35	72	16
Foot Length (cm)	0.17 ± 0.01	0.16 ± 0.01	0.17 ± 0.01
Foot Width (cm)	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01
Foot Area (cm^2)	1.30 ± 0.23	1.22 ± 0.20	1.34 ± 0.14
Hull Area (cm^2)	0.86 ± 0.12	0.81 ± 0.12	0.86 ± 0.09
LOP_Dev_Angle	0.28 ± 2.23	0.30 ± 2.50	0.40 ± 1.98
BOS Area (cm^2)	669.25 ± 103.48	622.74 ± 108.20	683.97 ± 94.05
Toe Angle	4.93 ± 3.33	5.97 ± 4.09	5.90 ± 3.61
Step Length (cm)	63.40 ± 7.05	57.48 ± 10.35	66.92 ± 6.29
Step Width (cm)	65.07 ±6.65	59.50 ± 9.57	68.09 ± 6.19
Stride Length (cm)	126.09 ± 14.11	114.31 ± 19.99	133.32 ± 12.72
Stride Width (cm)	13.43 ± 4.57	13.93 ± 4.25	11.51 ± 3.44
Base Width (cm)	11.97 ± 4.30	12.24 ± 4.27	9.69 ± 3.36
Step Time (sec)	0.62 ± 0.48	0.63 ± 0.32	0.64 ± 0.16
Stride Time (sec)	1.25 ± 0.67	1.28 ± 0.50	1.30 ± 0.29
Step Velocity (cm/sec)	109.52 ± 21.60	96.61 ± 25.42	109.56 ± 21.60
Stride Velocity (cm/sec)	106.61 ± 21.08	94.41 ± 24.48	106.66 ± 21.11
Single Support Time (sec)	0.41 ± 0.06	0.42 ± 0.25	0.45 ± 0.13
Double Support Time (sec)	0.18 ± 0.05	0.21 ± 0.23	0.19 ± 0.05
Stance Time (sec)	0.78 ± 0.15	0.84 ± 0.47	0.82 ± 0.21

Table 4.1 shows the mean values of the gait parameters for all participants. Participants were divided into three groups: mild patients, moderate patients, and healthy controls. Gait data related to body size, such as foot length, foot width, foot area, and hull area, were first normalized according to the participants' height and then subjected to statistical analysis. Foot length, foot width, foot area, hull area and LOP_Dev_Angle showed little difference among the three groups. For the BOS area, step/stride length, width and base width, mild patients had a value that fell between that of moderate patients and healthy controls. While mild patients had the lowest toe angle, moderate patients had the lowest values for both step stride velocity. Healthy controls had a much lower stride width value than the other groups.

4.2 MS vs. Healthy Control Classification

4.2.1 Feature Selection

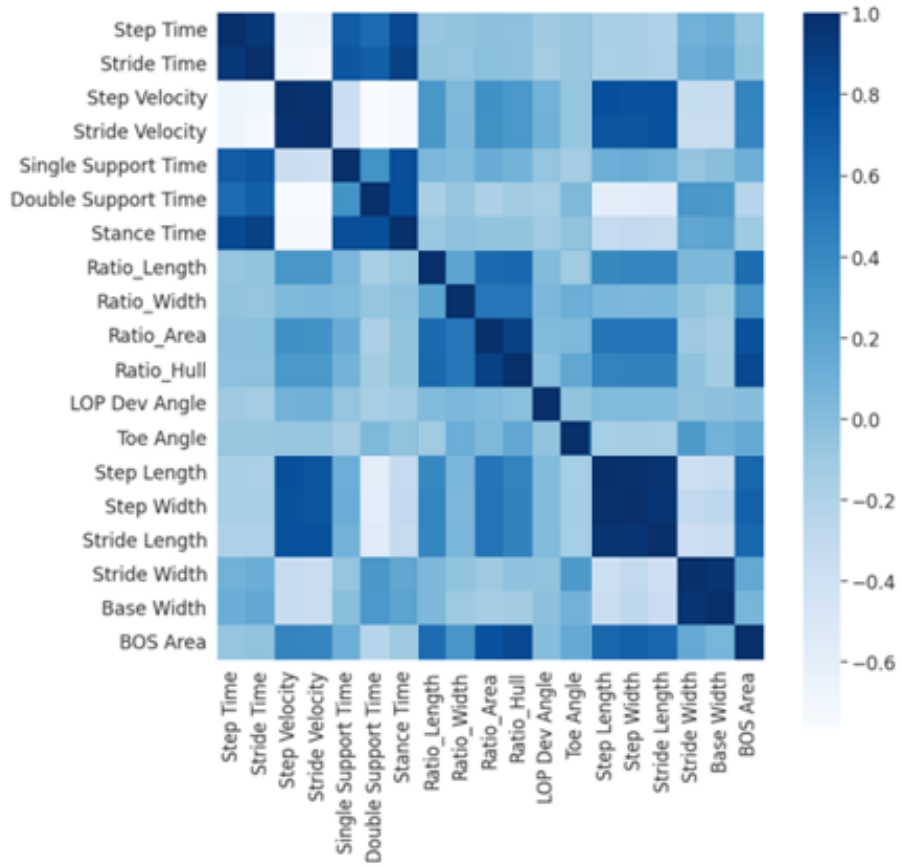

Figure 4.1 Heatmap for feature correlations for MS vs. healthy control classification. Heatmap regions that are increasingly dark show higher correlations.

Correlation analysis of all the numerical features was performed to determine their independence, as shown in Figure 4.1. The heatmap displays a strong positive correlation between the ‘step’ and ‘stride’ parameters (r >0.8), the base width and stride width. Stride time, stride velocity, and base width were excluded from further analysis to reduce interdependence among the features.

After removing the highly correlated features, feature selection was performed on both the standard and augmented feature sets to determine the optimal size of each set.

4.2.1.1 Selected Features

The feature selection method used for the MS vs. healthy control classification was ANOVA-SVM because the model performance was better than that of RFECV. The standard set was initialized with 10 numerical input features and one categorical feature (Table 4.2). The ANOVA-SVM suggested an optimal subset of nine features: step velocity, single support time, double support time, stance time, signed toe angle, step length, step width, stride length and stride width. After numerical feature selection was completed, the categorical feature, foot type, was reintroduced, resulting in the final standard set.

The augmented set was initialized with 16 numerical and 2 categorical features. The ANOVA-SVM selected 15 numerical features: step velocity, single support time, double support time, foot length, foot width, foot area, hull area, unsigned toe angle, step length, step width, stride length, stride width, and BOS area. The features step time, stance time and LOP angle were dropped. Once the processing of the numerical features was completed, the categorical features, foot type and toe direction, were reintroduced to the final augmented set. The original and final features are displayed in Table 4.2.

Table 4.2 Features for MS vs. healthy control classification.

Feature set	**Original Features Set**	**Final Features Set**
Standard set	Step time	
	Step velocity	Step velocity
	Single support time	Single support time
	Double support time	Double support time
	Stance time	Stance time
	Toe angle (signed)	Toe angle (signed)
	Step length	Step length
	Step width	Step width
	Stride length	Stride length
	Stride width	Stride width
	Foot type	Foot type

Augmented set	Step time	
	Step velocity	Step velocity
	Single support time	Single support time
	Double support time	Double support time
	Stance time	
	Toe angle (unsigned)	Toe angle(unsigned)
	Step length	Step length
	Step width	Step width
	Stride length	Stride length
	Stride width	Stride width
	Foot type	Foot type
	Toe direction (in/out)	Toe direction (in/out)
	Hull area	Hull area
	BOS area	BOS area
	LOP deviation angle	
	Foot length	Foot length
	Foot width	Foot width
	Foot area	Foot area

4.2.2 Feature Importance

The importance of each gait feature in differentiating classes was calculated. The average F-score from ANOVA-SVM (Tables 4.3 and 4.4), for feature correlation of LR, feature importance of XGB, and SHAP were tested. As SHAP is not supported by SVM when its kernel parameter is set as a radial basis function, it was calculated only for LR and XGB. The absolute value of a feature indicates the importance of the feature. The higher the absolute value, the more important the feature is.

Table 4.3 Average F- score for optimal standard set features

Features	**Feature Score**
Stride Length	974.54
Step Length	924.42
Step Width	862.72
Stride Width	330.42
Step Velocity	243.75
Single Support Time	37.58
Double Support Time	25.77
Toe Angle Signed	4.89

Stance Time	4.10

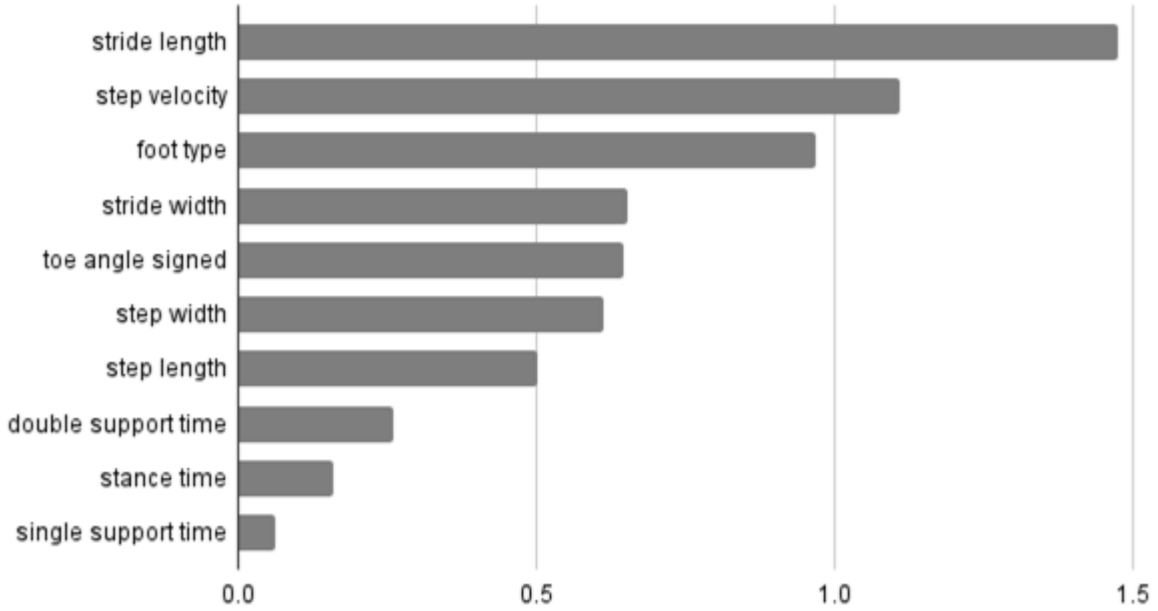

Figure 4.2 Absolute value of feature coefficient of LR using the standard set (MS vs. healthy control classification)

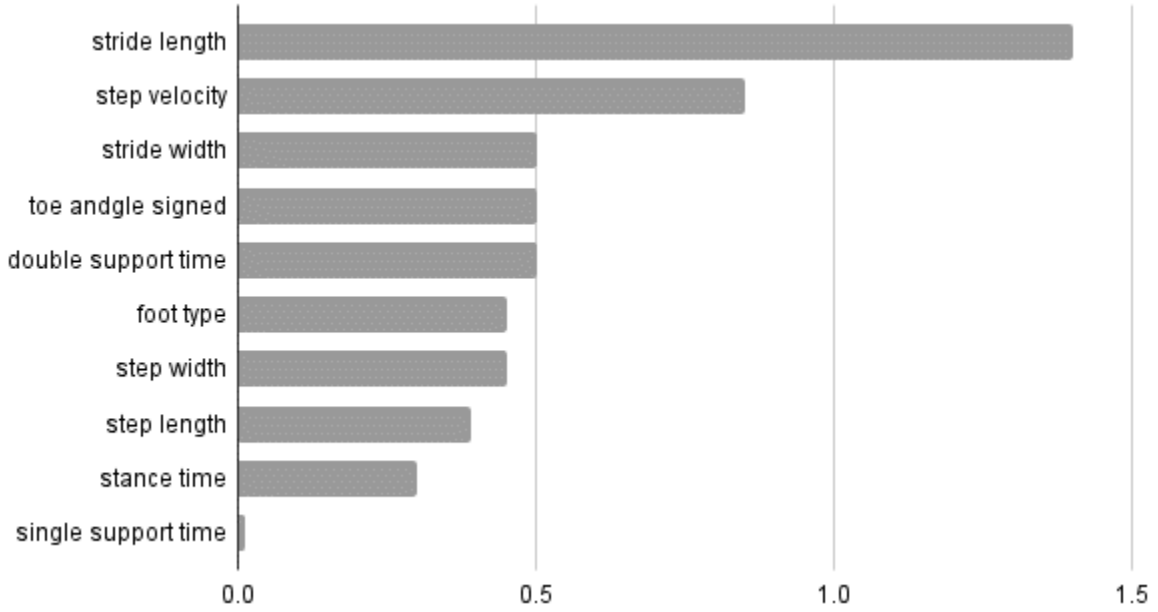

Figure 4.3 Mean SHAP value of LR using the standard set (MS vs. healthy control classification)

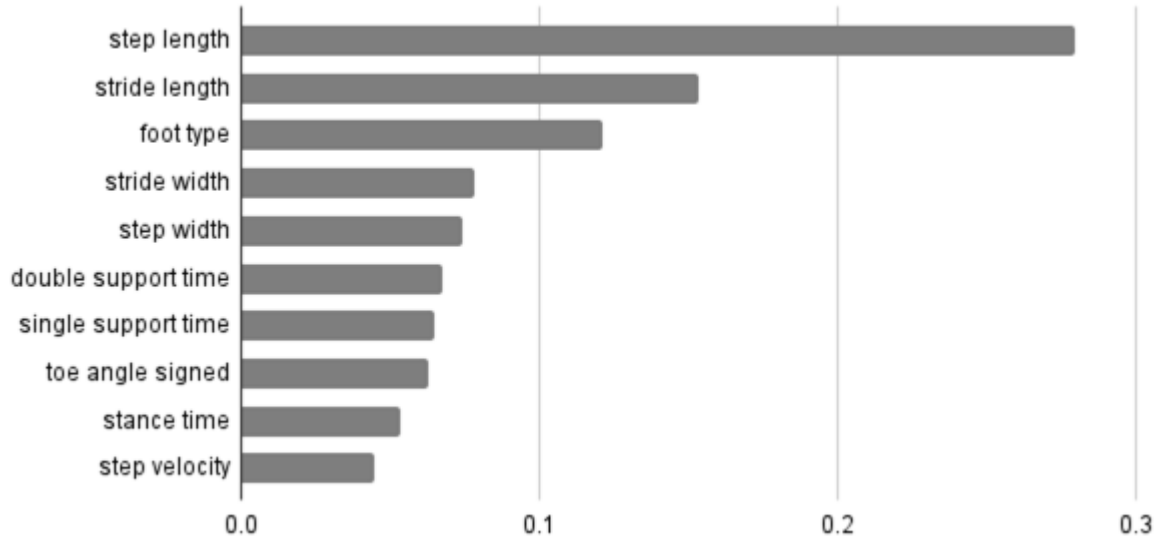

Figure 4.4 Feature importance of XGB using the standard set (MS vs. healthy control classification)

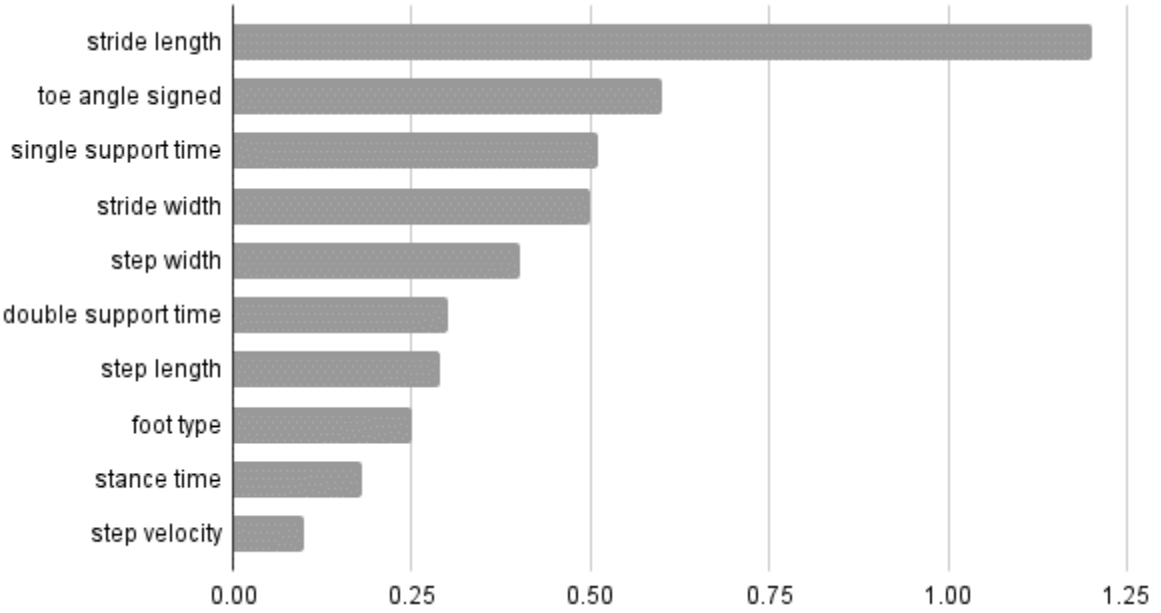

Figure 4.5 Mean SHAP value of XGB using the standard set (MS vs. healthy control classification).

The figures above present the importance of each feature in the final standard set of this experiment. The results from all methods indicate that stride length was the most important feature for prediction. Table 4.3 presented the average F-score of each selected feature from the ANOVA-SVM. The stride length, step length and step width had the highest F-scores. Figures 4.2 and 4.3 show that stride length and step velocity are the most important features of the LR

classifier. Figures 4.4 and 4.5 display the importance of the features of the XGB classifier. The toe angle was evaluated as the second most important feature in the prediction using the SHAP value. However, the second most important feature calculated for XGB was stride length, followed by step length.

Table 4.4 Average F-score for optimal augmented set features

Features	Feature Score
Stride Length	977.28
Step Length	934.28
Step Width	874.32
Foot Area	374.92
BOS Area	281.11
Stride Width	265.59
Step Velocity	250.78
Hull Area	222.84
Foot Length	91.88
Single Support Time	67.70
Double Support Time	29.05
Foot Width	19.32
Toe Angle Unsigned	2.37

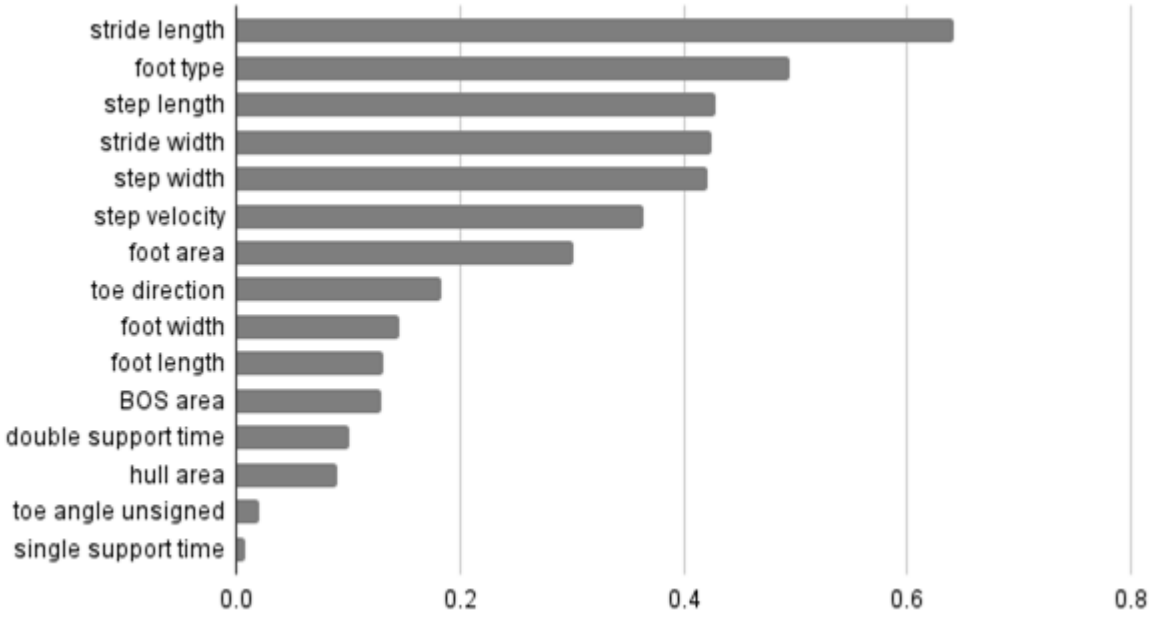

Figure 4.6 Absolute value of feature coefficient of LR using the augmented set (MS vs. healthy control classification)

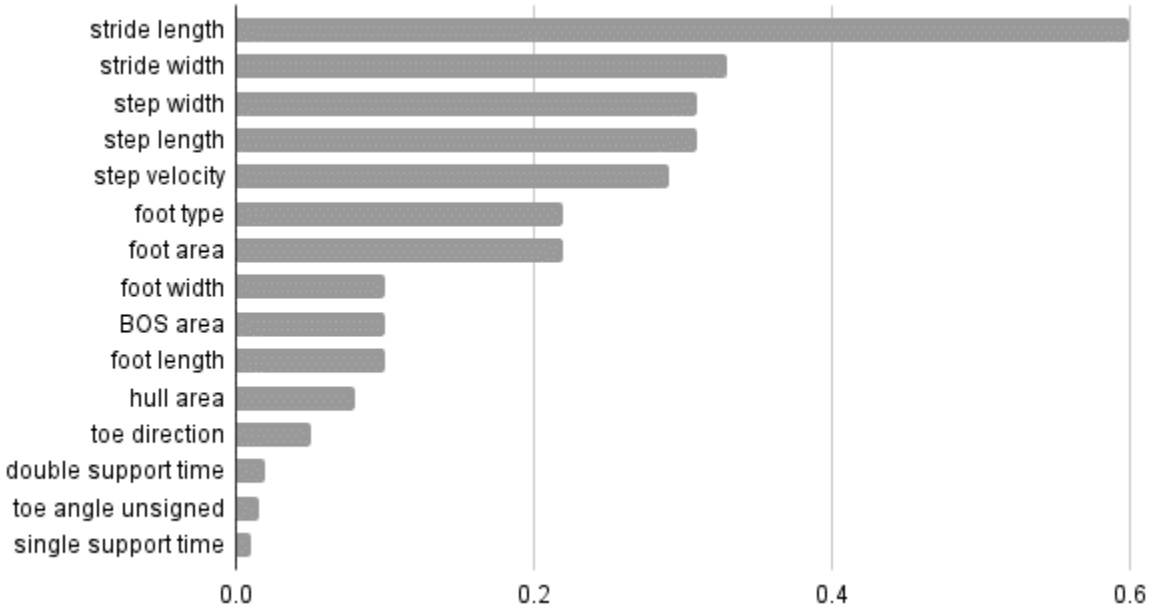

Figure 4.7 Mean SHAP value of LR using the augmented set (MS vs. healthy control classification)

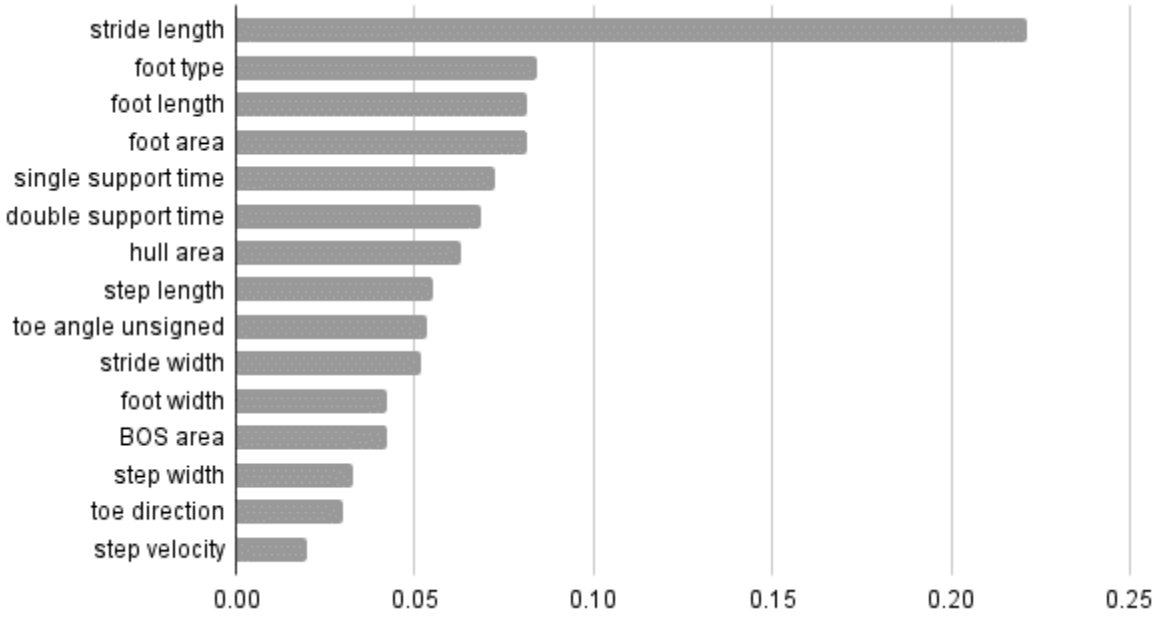

Figure 4.8 Feature importance value of XGB using the augmented set (MS vs. healthy control classification)

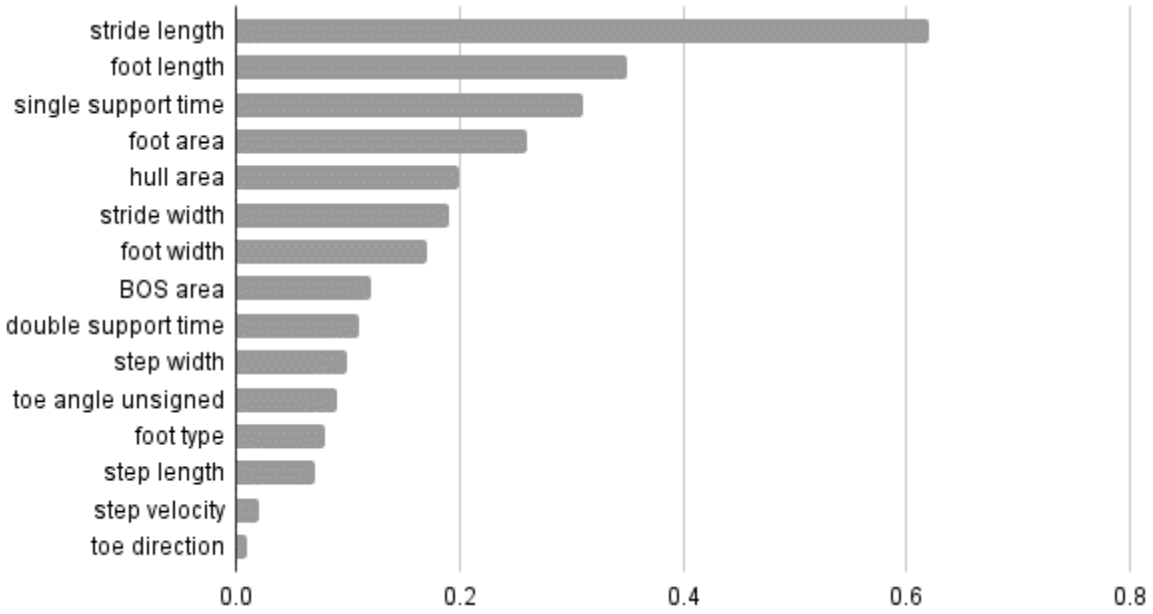

Figure 4.9 Mean SHAP value of XGB using the augmented set (MS vs. healthy control classification)

Table 4.4 indicates that, for the augmented set, the most important features selected by the ANOVA-SVM were the stride length, step length and stride width, followed by new features, foot area and BOS area. Figures 4.6 and 4.7 display the importance of the features of the LR classifier. The most important feature was stride length. The second most important feature was foot type, as evaluated by the feature coefficient of LR; SHAP evaluated the stride width as the second most important feature for LR. Figures 4.8 and 4.9 present the importance of the features of the XGB classifier. While stride length is the most critical feature of XGB, this classifier also considered the new features of foot length and foot area, which contributed significantly to the model prediction result.

4.2.3 Prediction Results

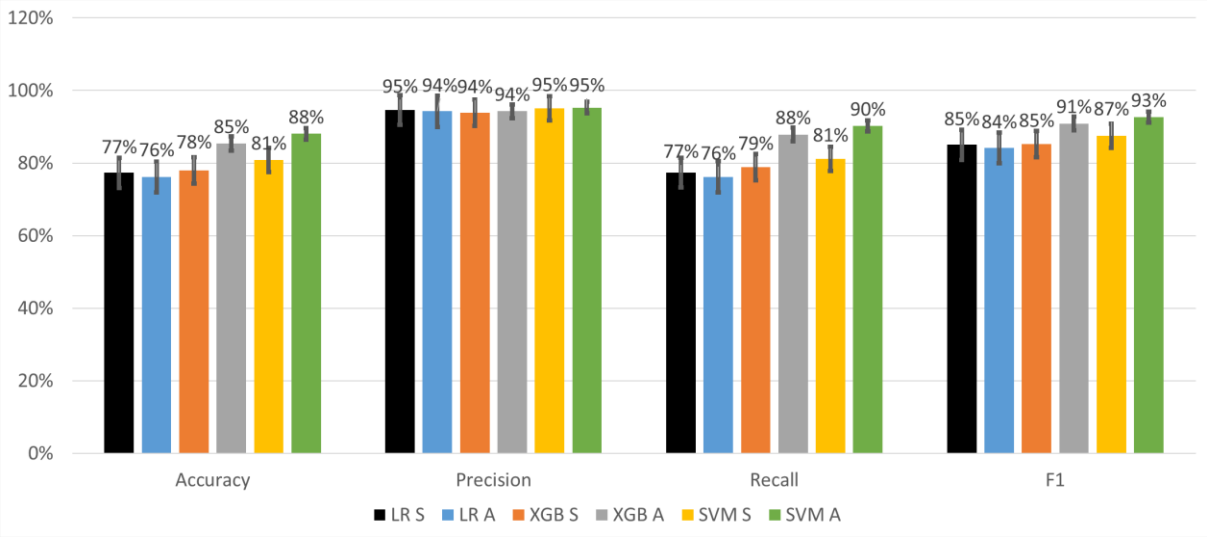

Figure 4.10 Accuracy, precision, recall and F1 score for each model of MS vs. healthy control classification. S refers to the standard set, and A refers to the augmented set.

When the standard feature set was used, the highest accuracy of 81% (using SVM), the precision of 94% (SVM and LR), recall of 81% (SVM) and F1 score of 87% (SVM) were achieved. The details of these comparisons are presented in Figure 4.10.

The improvements measured across all metrics using the augmented feature set are also worth noting. The inclusion of the extra features increased the accuracy by 7%, recall by 9%, and F1 score by 6% from both the XGB and SVM models. Note that precision was not improved because of the imbalance in the testing dataset, where the number of false positives was relatively small compared to that of true positives. The strongest improvement in the scores was obtained using SVM.

In addition to the scoring metrics, the area underneath the receiver operating characteristic (AUROC) and the area underneath the precision-recall curves (AUPRC) was also used to determine a classifier's overall effectiveness.

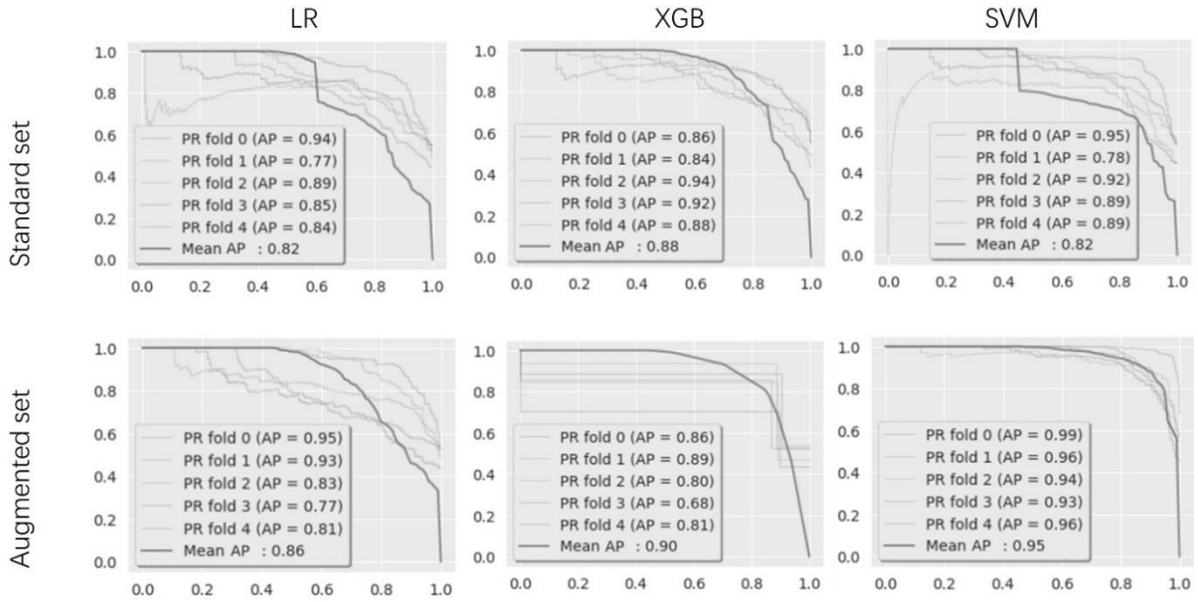

Figure 4.11 PRC curves for LR, XGB and SVM. AP refers to average precision.

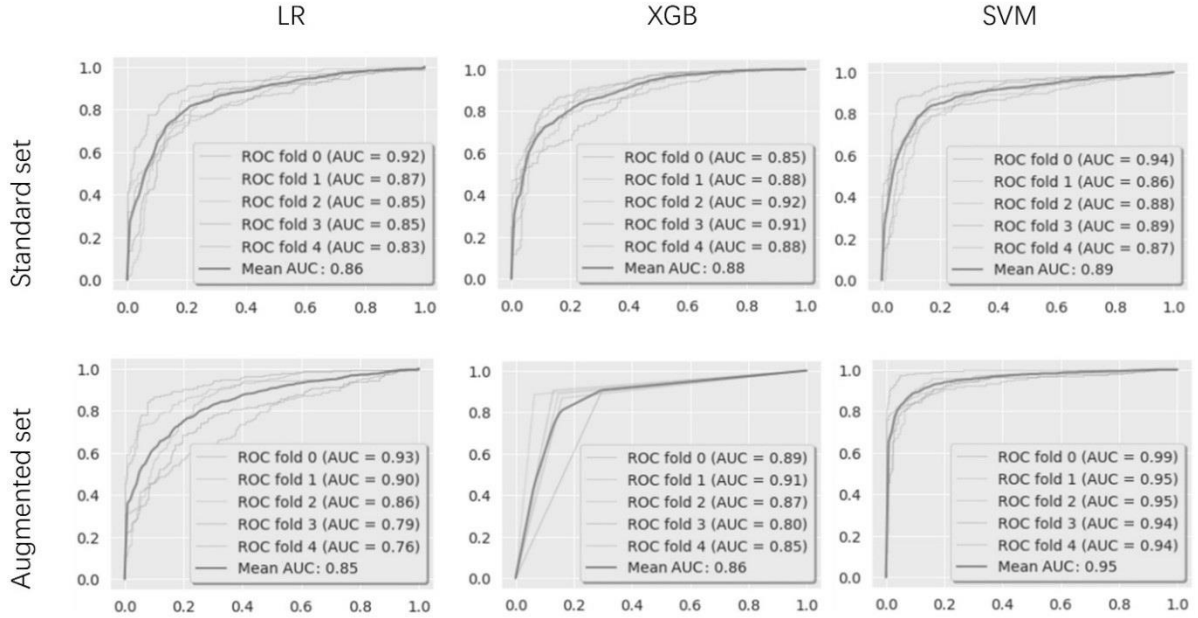

Figure 4.12 ROC curves for LR, XGB and SVM. AUC refers to the area under the curves.

According to Figures 4.11 and 4.12, when studying the standard feature set, the best baseline of AUROC was 0.88 (XGB), and for AUPRC, it was 0.89 (SVM). The variance measured between all classifiers was low with these scoring metrics, resulting in similar scores for all models

AUPRC and AUROC metrics were compared for the augmented feature set, as well. When using the augmented set, the AUROC of LR and XGB did not improve; however, the AUROC increased when using SVM and AUPRC improved for all models.

4.3 Mild MS vs. Moderate MS Classifications

4.3.1 Feature Selection

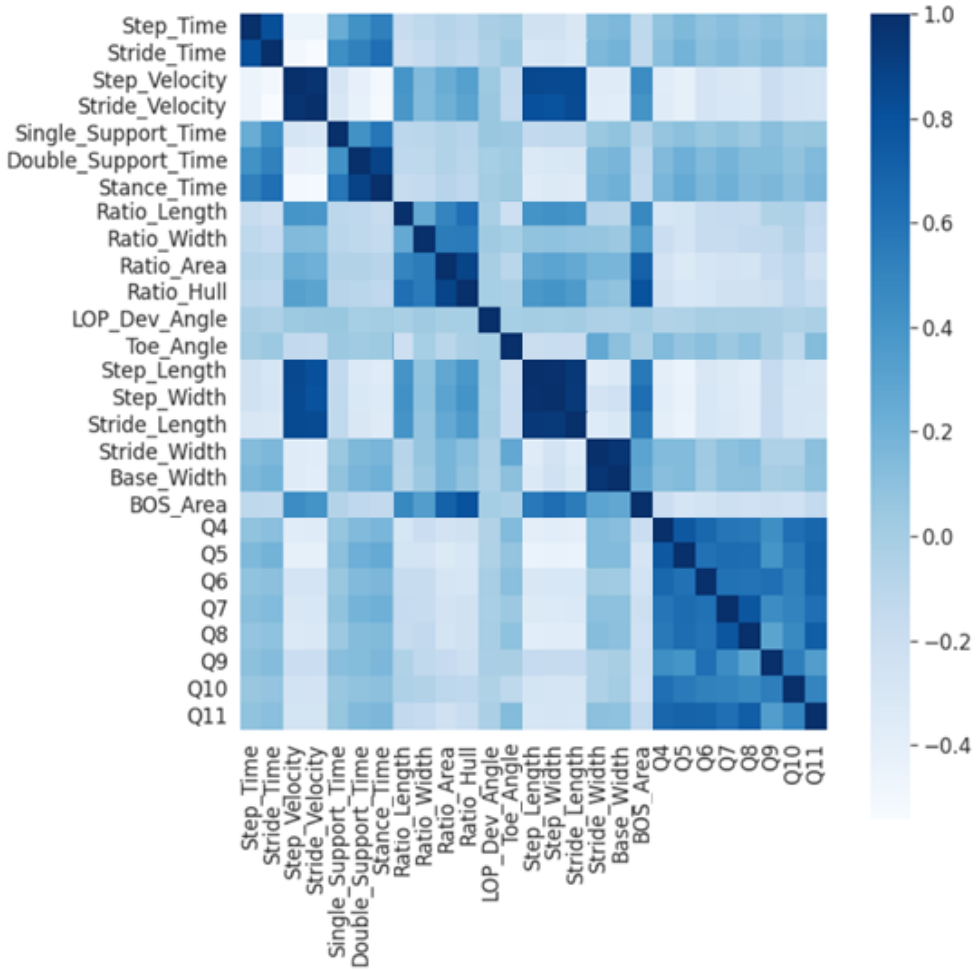

Figure 4.13 Heatmap for feature correlations for mild MS vs. moderate MS patient classification. Heatmap regions that are increasingly dark (light) show areas of higher (lower) correlations. Q4-Q11 represent the MSIS-29 questions 4 to 11.

For the mild/moderate patient task, correlations between the MSIS-29 subset score and other features were also calculated. The heatmap for the mild/moderate patient experiment showed that none of the features were significantly correlated with any of the questions. The correlations between features were almost the same as those in Experiment 1.

4.3.1.1 Selected Features

ANOVA-SVM provided better results than RFECV in this experiment. For both the standard and augmented sets, ANOVA-SVM determined that all features could be selected, and that all contributed to model prediction (Table 4.5).

Table 4.5 Features for mild MS vs. moderate MS patient classification.

Feature set	**Features**
Standard set	Step time Step velocity Stride velocity Single support time Double support time Stance time Toe angle (signed) Step length Step width Stride length Stride width Foot type
Augmented set	Step time Step velocity Single support time Double support time Stance time Foot length Foot width Foot area Hull area LOP deviation angle Toe angle (unsigned) Step length Step width Stride length

	Stride width BOS area Foot type

4.3.2 Feature Importance

Feature importance was also calculated for mild/moderate patient classification. The average F-score value for ANOVA-SVM, feature coefficient for LR, feature importance for XGB and mean SHAP value for both LR and XGB are presented below.

Table 4.6 Average F- score for optimal standard set features

Features	Feature Score
step time	271.74
stride width	193.39
step width	108.83
step length	87.43
step velocity	64.30
single support time	63.04
stance time	23.92
toe angle (signed)	16.15
stride length	12.39
double support time	12.16

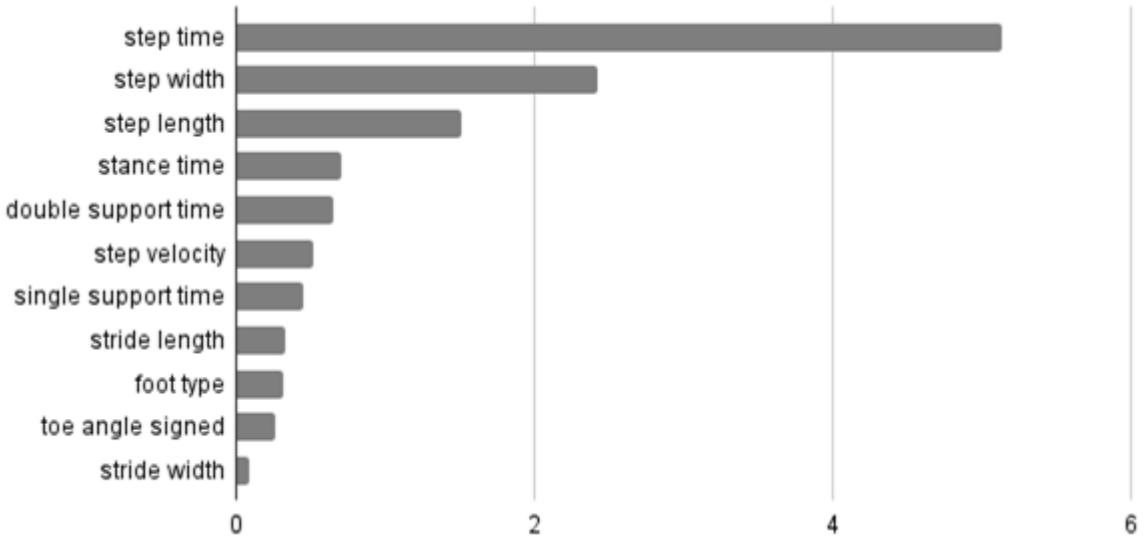

Figure 4.14 Absolute value of feature coefficient of LR using the standard set (mild MS vs. moderate MS patient classification)

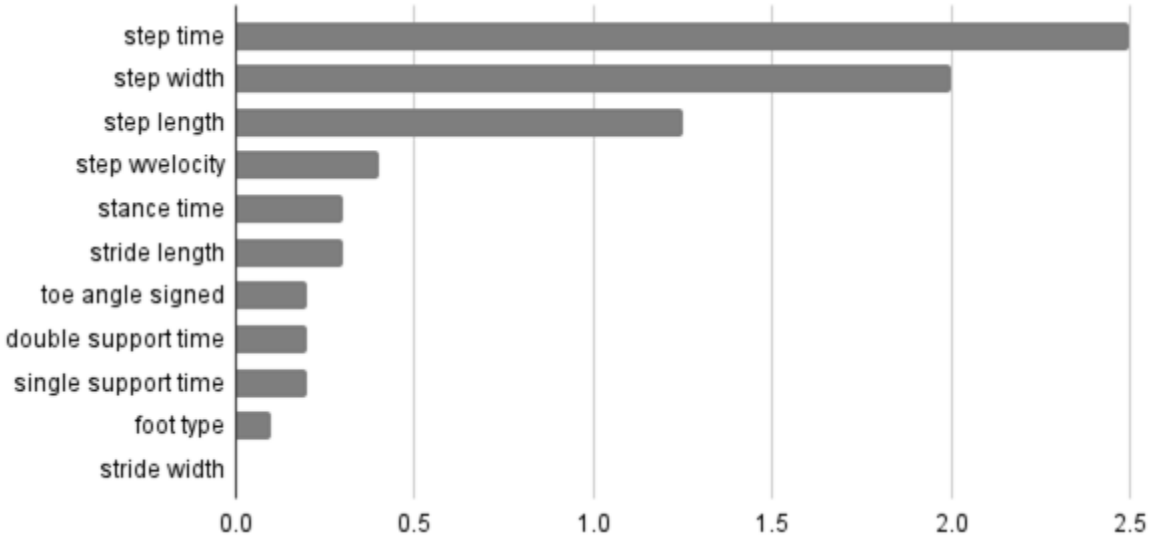

Figure 4.15 Mean SHAP value of LR using the standard set (mild MS vs. moderate MS patient classification)

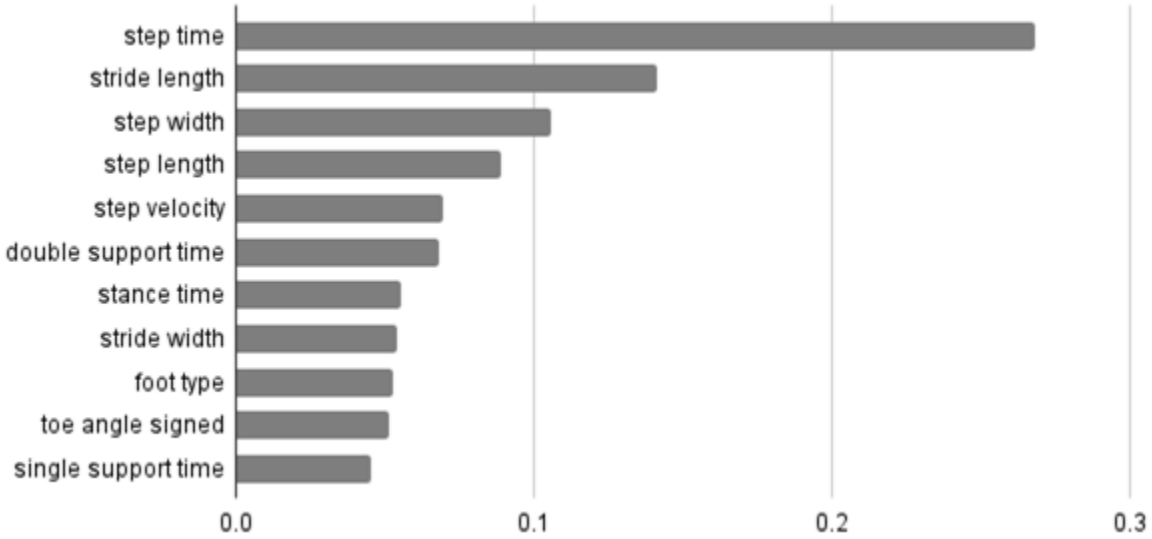

Figure 4.16 Feature importance of XGB using the standard set (mild MS vs. moderate MS patient classification)

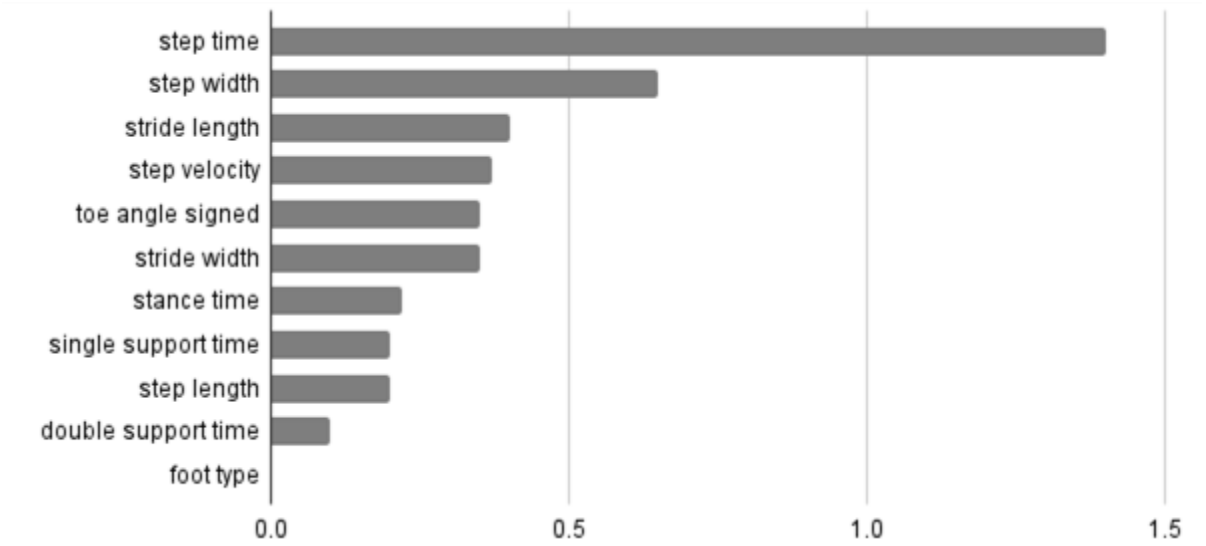

Figure 4.17 Mean SHAP value of XGB using the standard set (mild MS vs. moderate MS patient classification)

Table 4.6 shows that step time, stride width, and step width are the top three important features evaluated by ANOVA-SVM. Figures 4.14 to 4.17 also demonstrate that step time and step width were the two features that contributed the most to model prediction. According to the feature importance and SHAP value, stride length was important to the XGB model result.

Table 4.7 Average F- score for optimal augmented set features

Features	**Feature Score**
step width	561.98
step length	504.57
BOS area	416.50
hull area	402.72
stride length	354.94
foot area	339.31
foot length	261.41
step velocity	181.54
foot width	60.07
toe angle	46.48
double support time	39.89
stance time	35.75
step time	35.70
LOP Dev angle	10.66
single support time	4.90
stride width	1.41

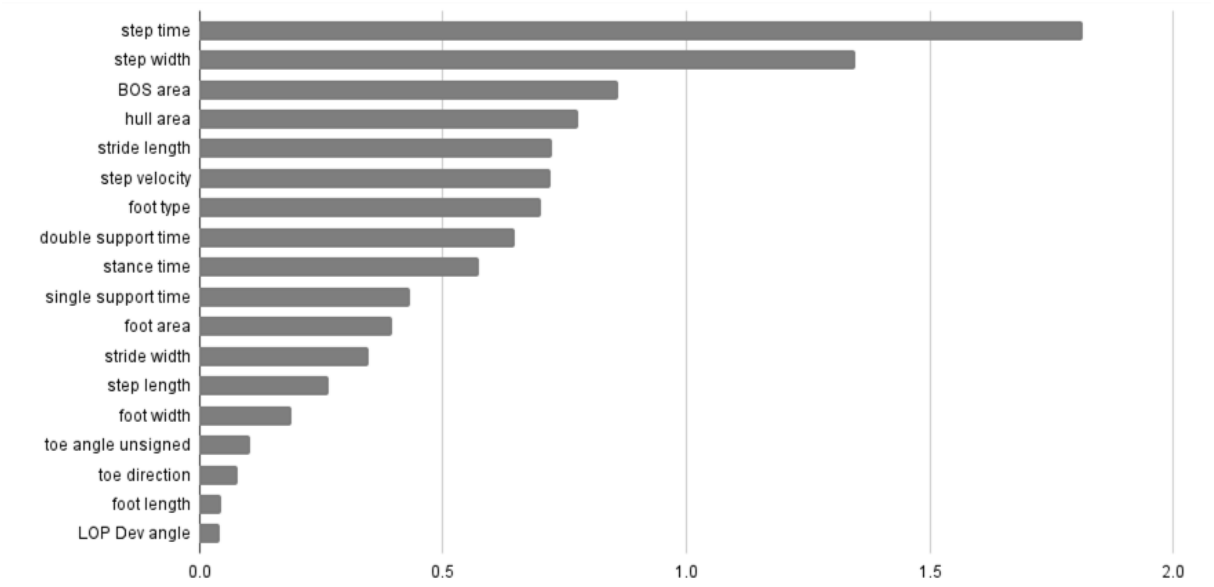

Figure 4.18 Absolute value of feature coefficient of LR using the augmented set (mild MS vs. moderate MS patient classification)

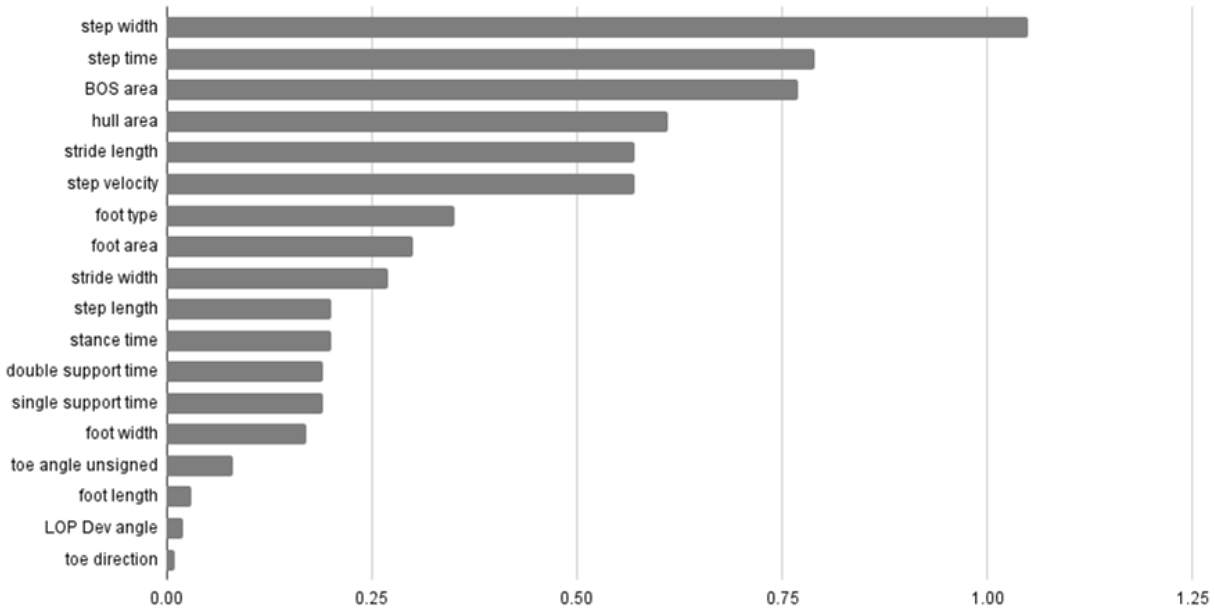

Figure 4.19 Mean SHAP value of LR using the augmented set (mild MS vs. moderate MS patient classification)

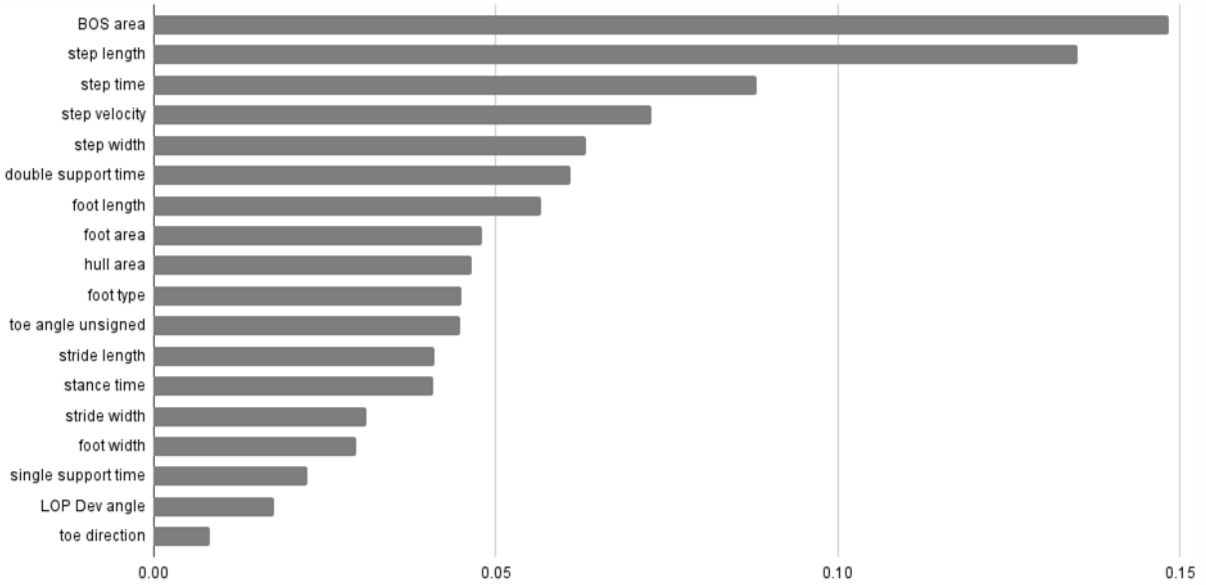

Figure 4.20 Feature importance of XGB using the augmented set (mild MS vs. moderate MS patient classification)

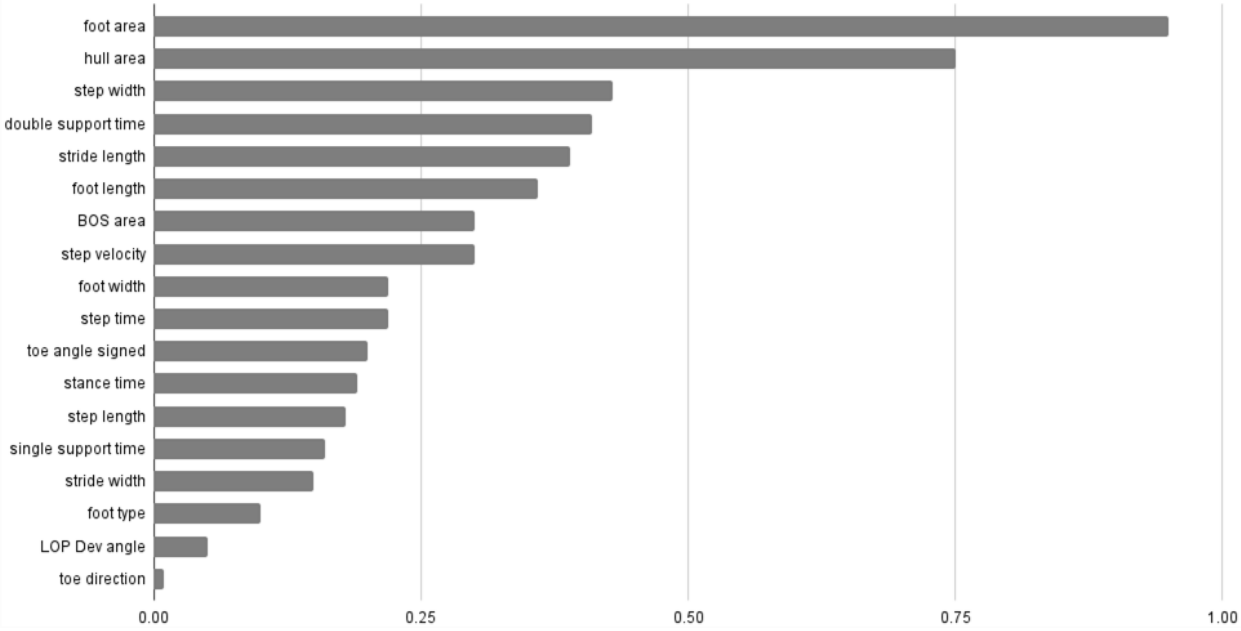

Figure 4.21 Mean SHAP value of XGB using the augmented set (mild MS vs. moderate MS patient classification)

Table 4.7 indicates that step width, step length, BOS area, and hull area are the top four most important features for the augmented set. As step width and area-related features are still of great importance in LR prediction, as presented in Figures 4.18 and 4.19, the feature step time is of importance in the classification. For the XGB classifier, Figures 4.20 and 4.21 show

that area-related features, BOS area, foot area and hull area, are the most important features. The step time and step width are less important than the area-related features.

4.3.3 Prediction Results

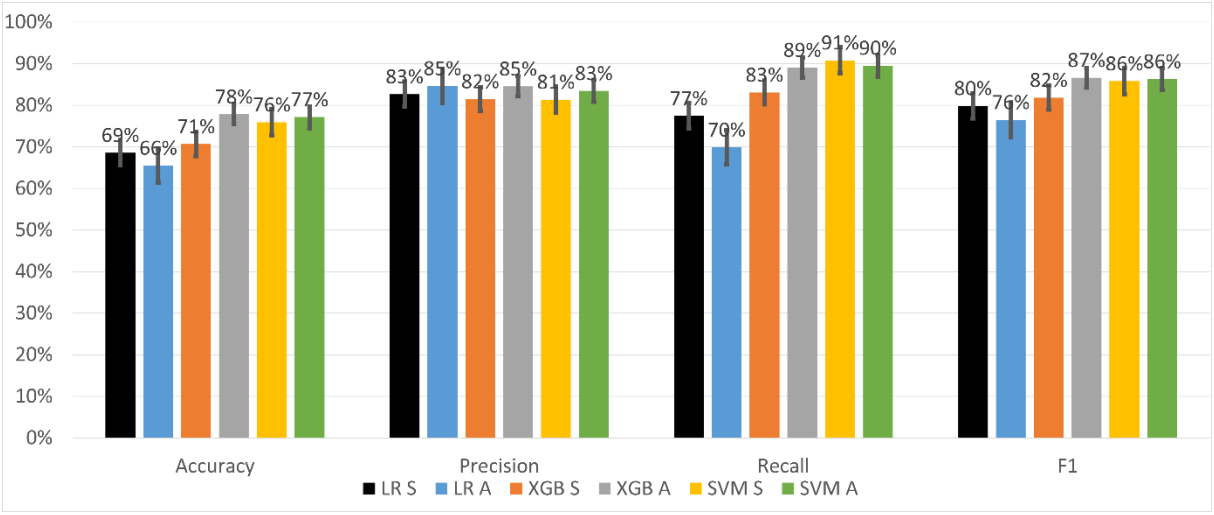

Figure 4.22 Accuracy, precision, recall and F1 score for each model of mild MS vs. moderate MS patient classification. S refers to the standard set, and A refers to the augmented set.

Accuracy, precision, recall and F1 score were calculated for the mild MS vs. moderate MS patient classification (Figure 4.22). For the standard set, the best results were accuracy of 76% (SVM), precision of 83% (LR), recall of 91% (SVM), and F1 score of 86% (SVM). When the augmented set was introduced to the models, the model performance slightly increased in accuracy by 2% (XGB), precision by 2% (LR), and F1 score by 1% for XGB.

The AUROC and AUPRC were also printed for both feature sets. Figures 4.23 and 4.24 present the details of AUROC and AUPRC, respectively.

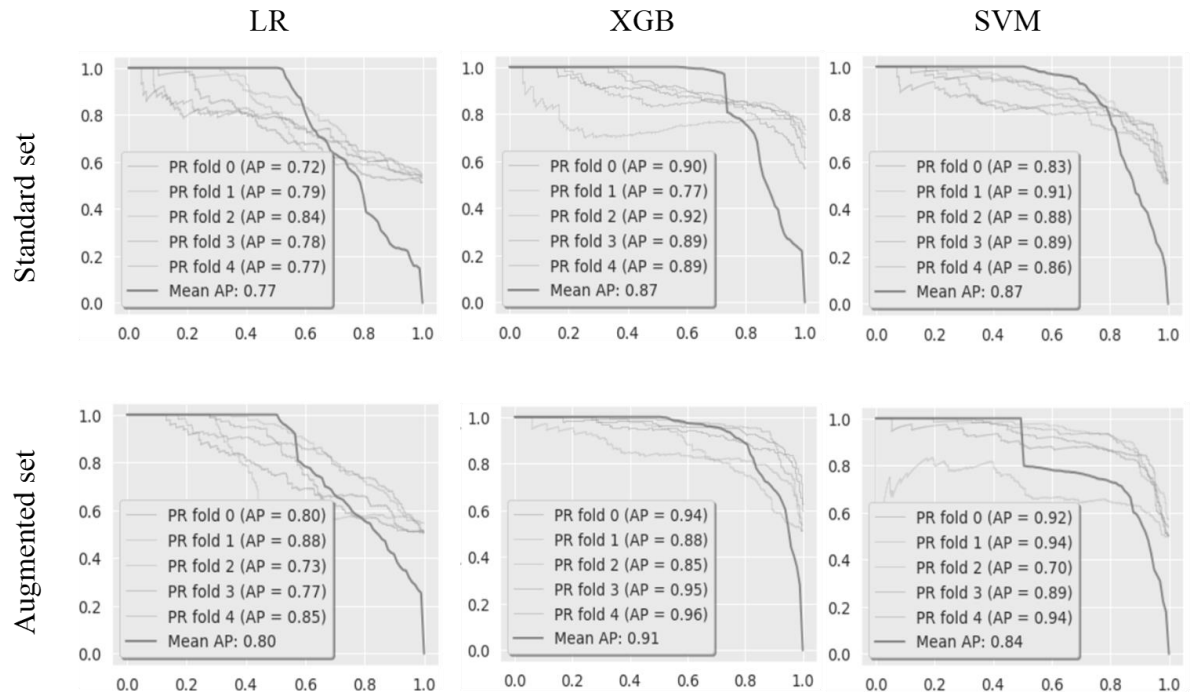

Figure 4.23 PRC curves for LR, XGB and SVM. AP refers to average precision

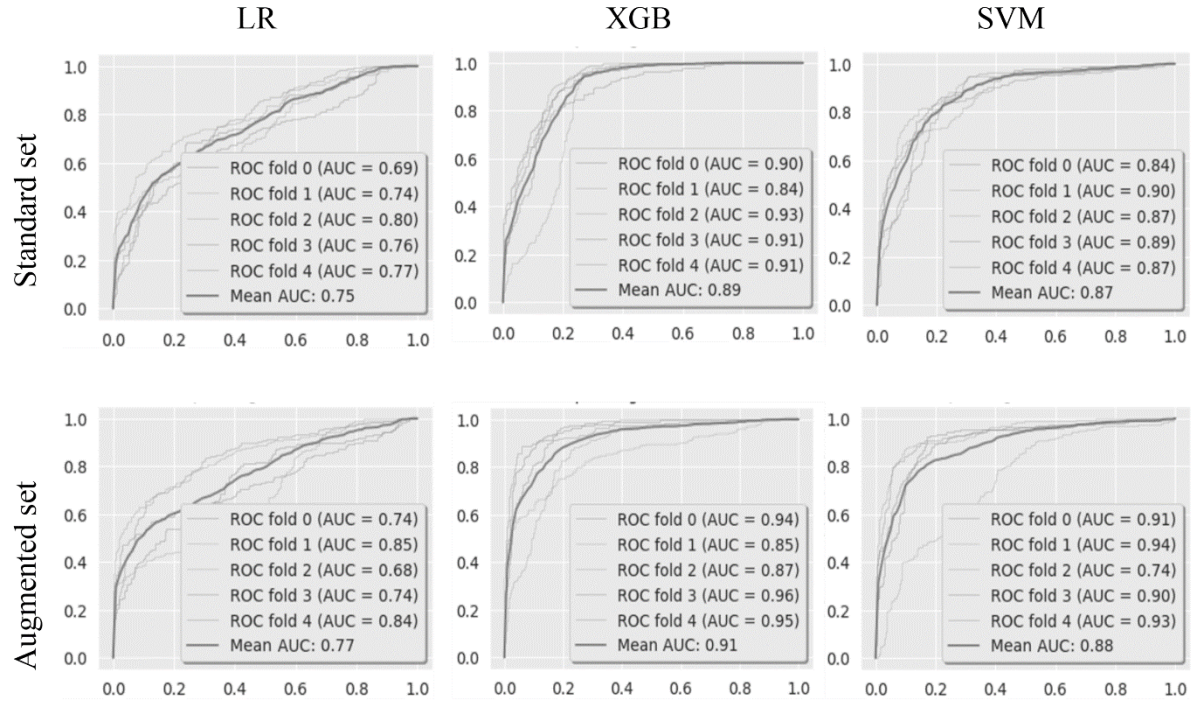

Figure 4.24 ROC curves for LR, XGB and SVM. AUC refers to the area under curves.

The best baseline AUPRC was achieved with 0.87 (SVM and XGB) and the best baseline AUROC was 0.89 (XGB). The use of the augmented set slightly improved the AUPRC and AUROC of the LR and XGB.

5 Discussion and Conclusion

This chapter summarizes the findings in this study and discusses possible future improvements.

5.1 Discussion

The research question, whether gait features calculated from the raw walkway sensor data can be used to separate healthy controls from MS patients or to classify patients with different gait disability severities, has been addressed by the results presented in this thesis. The use of the standard set for Experiment 1, MS vs. Healthy control classification, provided a prediction accuracy of 81%, and including the newly created or unutilized features generated from the raw walkway data improved the model performance by 7%. For Experiment 2, mild MS vs. moderate MS patient classification, using the standard set provided a baseline accuracy of 76%, and the augmented set improved the accuracy by 2%.

The main difference between this thesis and previous works [46], [48] is that our features were derived from raw sensor data. This study focused on these new features to prove that the raw walkway sensor data can provide clinicians with unique features for gait analysis. In contrast to wearable sensors and visual systems, walkway sensors can detect and record a stream of continuous data that reflects the relationship between steps. The walkway makes it possible to create unique features that are rarely used in gait studies, such as the line of

progression deviation angle, toe angle, and base of the support area. It was demonstrated that these new features improved the prediction performance.

5.1.1 Data Balancing Methods

Two data balancing methods were used in this thesis: SMOTE and ADASYN. SMOTE yielded better results for the patient/control separation experiment, whereas ADASYN yielded higher results for patient severity prediction. SMOTE and ADASYN execute similar algorithms to create new samples. They both create new samples via interpolation. However, these two methods differ in the approach taken for the new sample generation. SMOTE chooses a sample and the sample's k-nearest neighbors from the minority class generate new samples from line segments that connect these samples, and new samples are selected from the line. This process is repeated until the data of the two classes are balanced. ADASYN creates new data next to the original samples that the k-nearest neighbors incorrectly classify. Therefore, the possibility of obtaining a correct prediction on these samples is increased because the model learns more from them. Such a method may be useful in situations where the labeled classes have data points close to each other, such as the dataset of mild MS vs. moderate MS patient classification in this thesis.

5.1.2 Feature Selection Methods

Two feature selection methods (RFECV and ANOVA-SVM) were tested in this thesis and the results for both methods were very similar, with ANOVA-SVM providing slightly better results. The ANOVA-SVM method calculates the correlation between the features and targets,

and only the features that are highly correlated with the targets are selected before proceeding. RFECV determines the final feature set according to the importance of each feature based on their cross-validation scores. Both methods can choose the candidate features that would maximize model prediction. RFECV takes much longer than ANOVA-SVM to complete the feature selection process. Therefore, when dealing with very large feature sets, ANOVA-SVM is a better choice. Because RFECV requires the feature coefficient or feature importance attributes of the model to determine which features to choose, and thus, the use of this method may be limited.

5.1.3 Feature Importance

The sets of features determined as the most important for the two classification experiments were different. When using the standard set to classify MS patients from healthy controls, all feature importance calculation methods valued stride length the most. The feature step velocity was considered the second most important feature for the LR classifier. The signed toe angle was selected by LR, and the average SHAP value of the features of the XGB was important for distinguishing patients from healthy controls. For the augmented set, stride length was still determined to be the most important feature, indicating that stride length was found to be very useful for discriminating between MS patients and healthy controls. This can be a pointer for clinicians or therapists when studying the rehabilitation effect in MS patients. In addition to the stride/step length/width features, foot area and foot length were also important for both LR and XGB as they contributed significantly to the classification result. Foot area and foot length were

considered as new features in this study, whereas they have been usually ignored in other gait studies. However, the results of our study indicate that the shape of the foot of MS patients is important for classification, indicating that it may be affected by the disease or the ongoing gait disability.

In contrast to MS patients vs. healthy control classification, stride length was not the most important feature of the classifiers when using the standard set for classifying mild patients from moderate patients. Instead, step time and step width were evaluated as the two most important features of classifiers. This may indicate that patients with different severity levels might have similar stride lengths, whereas their step time and step width vary. As reported by Brach et al. in 2005, people with either low or high step width were more likely to report a fall [87], and step width can also be an indicator in identifying patients with potential gait disability. When using the augmented set, LR still valued step width and step time as the most important features. However, area-related features were preferred by XGB. When determining the disease severity, the BOS area, foot area and hull area were of higher importance to the XGB classifier. It is worth noticing that, similar to Experiment 1, this result may indicate that the foot might be affected by the progression of the disease.

5.2 Conclusion

This thesis demonstrates that machine learning can be used to distinguish healthy controls from MS patients as well as classify patients' disability levels using only the raw data collected from an instrumented walkway system. Advances in computerized machine learning and

classification can easily handle complicated underlying sensor data, thus enabling researchers to create new gait measurements to detect gait issues automatically and rapidly.

This thesis has chosen to study gait by creating features from the raw underlying data instead of using gait measurements using gait analysis software. This allows the reconstruction of the standard gait parameters and the development of new features, such as the base of support area, line of progression deviation angle, hull area and toe direction, using unsupervised learning techniques like k-means. These standard and newly created parameters were then provided to the machine learning classifiers to determine the separability of the targets.

When trying to differentiate MS patients from healthy controls, the machine learning system discussed in this thesis achieved a base classification accuracy of 81% using only standard spatial and temporal gait parameters derived from the raw data. When these standard parameters were augmented with other custom parameters, the classification accuracy of the SVM was increased to 88%. Stride length and step width are two features that are highly recommended for MS patients and healthy controls classification studies. When classifying mild patients from moderate patients, the base accuracy score was 76% using the standard features; the inclusion of the augmented features increased the model performance by 2% for the XGB classifier. Stride time and step time are the two features that contribute the most to the classification of mild patients from moderate patients. At the same time, base of support area, foot area, and hull area functioned as the assistive features in the classification.

Both classification experiment results demonstrate that SVM and XGB models are suitable for analyzing raw gait data, and customizing new gait measurements is a worthwhile endeavor in increasing classification accuracy.

5.3 Future Work

The results obtained in this thesis are promising in identifying MS patients with gait-related dysfunction. Several improvements have been identified for future studies, which may further increase the usefulness of the results for gait researchers and clinicians.

The first involves pre-screening patients based on the MSIS-29 intake survey [70], [88]. The first experiment examined the separability of healthy controls and MS patients. Patients in this experiment answered gait-related MSIS-29 questions with at least one score equal to or higher than 3. This experiment showed that MS patients could be effectively classified from healthy controls. However, by including patients who report lower MSIS-29 scores, it may be possible, in future studies, to classify healthy controls from patients with milder symptoms of gait dysfunction [19], [20].

The second improvement could involve using pressure level data on top of the temporal and spatial data available from the walkway systems. Pressure data can be calculated according to the sensor record of the participants walking timeline and thus provide researchers and clinicians with more details on the way patients walk. This may enrich the dataset and is likely to be helpful in further enhancing classification accuracy. For instance, machine learning can be useful for mapping changes in specific types of gait impairments, such as those resulting

from hemiplegia or ataxia, with the help of pressure level data over time. Furthermore, using deep learning approaches to create features instead of using “hand-crafted” method can be considered a part of future work.

Finally, machine learning models can be improved using a larger training dataset. Previous studies have shown that machine learning technology combined with gait measurements can effectively distinguish patients with cognitive impairment levels [16]. Coordinating efforts between laboratories and research hospitals could result in a dataset of thousands of subjects, allowing machine learning models to train on a much richer set of underlying data, leading to more robust conclusions.

Bibliography

[1] D. S. Reich, C. F. Lucchinetti, and P. A. Calabresi, "Multiple sclerosis," *New England Journal of Medicine*, vol. 378, no. 2, pp. 169–180, 2018, doi: 10.1056/NEJMra1401483.

[2] T. M. S. I. F. (MSIF, "Atlas of MS 3 rd edition," vol. third edit, no. September, pp. 1–36, 2020, [Online]. Available: www.atlasofms.org.%0A

[3] E. A. Hakim *et al.*, "The social impact of multiple sclerosis - A study of 305 patients and their relatives," *Disability and Rehabilitation*, vol. 22, no. 6, pp. 288–293, 2000, doi: 10.1080/096382800296755.

[4] H. H. Scheinberg L, Holland N, Larocca N, Laitin P, Bennett A, "Multiple sclerosis; earning a living.," *N Y State J Med.N Y State J Med.*, vol. 80, 1980.

[5] N. G. Larocca, "Impact of Walking Impairment in Multiple Sclerosis Perspectives of Patients and Care Partners Conclusions: Difficulty walking is a common impairment in people with MS, with adverse effects on the QOL of people with MS and care partners of a person with MS," *Patient*, vol. 4, no. 3, pp. 189–201, 2011.

[6] C. Heesen, J. Böhm, C. Reich, J. Kasper, M. Goebel, and S. M. Gold, "Patient perception of bodily functions in multiple sclerosis: Gait and visual function are the most valuable," *Multiple Sclerosis*, vol. 14, no. 7, pp. 988–991, 2008, doi: 10.1177/1352458508088916.

[7] M. J. Socie, R. W. Motl, J. H. Pula, B. M. Sandroff, and J. J. Sosnoff, "Gait variability and disability in multiple sclerosis," *Gait and Posture*, vol. 38, no. 1, pp. 51–55, May 2013, doi: 10.1016/j.gaitpost.2012.10.012.

[8] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983, doi: 10.1212/wnl.33.11.1444.

[9] J. Hobart, D. Lamping, R. Fitzpatrick, A. Riazi, and A. Thompson, "The multiple sclerosis impact scale (MSIS-29) a new patient-based outcome measure," *Brain*, vol. 124, no. 5, pp. 962–973, 2001, doi: 10.1093/brain/124.5.962.

[10] F. Bethoux and S. Bennett, "Evaluating Walking in Patients with Multiple Sclerosis," *International Journal of MS Care*, vol. 13, no. 1, pp. 4–14, 2011, doi: 10.7224/1537-2073-13.1.4.

[11] J. Rueterbories, E. G. Spaich, B. Larsen, and O. K. Andersen, "Methods for gait event detection and analysis in ambulatory systems," *Medical Engineering and Physics*, vol. 32, no. 6. pp. 545–552, Jul. 2010. doi: 10.1016/j.medengphy.2010.03.007.

[12] A. R. Chaves, A. J. Devasahayam, M. Riemenschneider, R. W. Pretty, and M. Ploughman, "Walking Training Enhances Corticospinal Excitability in Progressive

Multiple Sclerosis—A Pilot Study," *Frontiers in Neurology*, vol. 11, no. June, pp. 1–15, 2020, doi: 10.3389/fneur.2020.00422.

[13] S. Czarnuch and M. Ploughman, "Automated gait analysis in people with multiple sclerosis using two unreferenced depth imaging sensors: Preliminary steps," *Proceedings of the 29th International Conference on Image and Vision Computing New Zealand, IVCNZ*, no. November, 2014, doi: 10.13140/2.1.2187.6481.

[14] T. Seel, J. Raisch, and T. Schauer, "IMU-based joint angle measurement for gait analysis," *Sensors (Switzerland)*, vol. 14, no. 4, pp. 6891–6909, 2014, doi: 10.3390/s140406891.

[15] M. Galli *et al.*, "Gait pattern in myotonic dystrophy (Steinert disease): A kinematic, kinetic and EMG evaluation using 3D gait analysis," *Journal of the Neurological Sciences*, vol. 314, no. 1–2, pp. 83–87, 2012, doi: 10.1016/j.jns.2011.10.026.

[16] P. H. Chen, C. W. Lien, W. C. Wu, L. S. Lee, and J. S. Shaw, "Gait-Based Machine Learning for Classifying Patients with Different Types of Mild Cognitive Impairment," *Journal of Medical Systems*, vol. 44, no. 6, pp. 1–7, 2020, doi: 10.1007/s10916-020-01578-7.

[17] B. E., B. D., and B. R., "Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease," *Applied Soft Computing Journal*, vol. 94, p. 106494, 2020, doi: 10.1016/j.asoc.2020.106494.

[18] K. Trentzsch *et al.*, "Using machine learning algorithms for identifying gait parameters suitable to evaluate subtle changes in gait in people with multiple sclerosis," *Brain Sciences*, vol. 11, no. 8, 2021, doi: 10.3390/brainsci11081049.

[19] J. J. Sosnoff, M. Weikert, D. Dlugonski, D. C. Smith, and R. W. Motl, "Quantifying gait impairment in multiple sclerosis using GAITRite^{TM} technology," *Gait and Posture*, vol. 34, no. 1, pp. 145–147, 2011, doi: 10.1016/j.gaitpost.2011.03.020.

[20] U. Givon, G. Zeilig, and A. Achiron, "Gait analysis in multiple sclerosis: Characterization of temporal-spatial parameters using GAITRite functional ambulation system," *Gait and Posture*, vol. 29, no. 1, pp. 138–142, 2009, doi: 10.1016/j.gaitpost.2008.07.011.

[21] A. Chen, M. C. Kirkland, K. P. Wadden, E. M. Wallack, and M. Ploughman, "Reliability of gait and dual-task measures in multiple sclerosis," *Gait and Posture*, vol. 78, pp. 19–25, May 2020, doi: 10.1016/j.gaitpost.2020.03.004.

[22] W. Hu *et al.*, "Machine learning classification of multiple sclerosis patients based on raw data from an instrumented walkway," *BioMedical Engineering OnLine*, vol. 21, no. 1, p. 21, Dec. 2022, doi: 10.1186/s12938-022-00992-x.

[23] J. M. Baker, "Gait Disorders," *American Journal of Medicine*, vol. 131, no. 6, pp. 602–607, 2018, doi: 10.1016/j.amjmed.2017.11.051.

[24] M. W. Whittle, *Gait analysis : an introduction*. Edinburgh: Butterworth-Heinemann, 2005.

[25] M. B. S. L. J. K. Loudon Janice; Swift, *The Clinical Orthopedic Assessment Guide*. Champaign, USA: Human Kinetics, 2013.

[26] A. D. Grant, "Gait Analysis: Normal and Pathological Function," *JAMA*, vol. 304, no. 8, p. 907, Aug. 2010, doi: 10.1001/jama.2010.1210.

[27] P. Decavel and Y. Sagawa, "Gait quantification in multiple sclerosis: A single-centre experience of systematic evaluation," *Neurophysiologie Clinique*, vol. 49, no. 2. Elsevier Masson SAS, pp. 165–171, Apr. 01, 2019. doi: 10.1016/j.neucli.2019.01.004.

[28] A. Miller, "Gait event detection using a multilayer neural network," *Gait and Posture*, vol. 29, no. 4, pp. 542–545, 2009, doi: 10.1016/j.gaitpost.2008.12.003.

[29] M. H. Cameron and J. M. Wagner, "Gait abnormalities in multiple sclerosis: Pathogenesis, evaluation, and advances in treatment," *Current Neurology and Neuroscience Reports*, vol. 11, no. 5, pp. 507–515, Oct. 2011, doi: 10.1007/s11910-011-0214-y.

[30] B. M. Meyer *et al.*, "Wearables and Deep Learning Classify Fall Risk from Gait in Multiple Sclerosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1824–1831, May 2021, doi: 10.1109/JBHI.2020.3025049.

[31] J. Han, H. S. Jeon, B. S. Jeon, and K. S. Park, "Gait detection from three dimensional acceleration signals of ankles in patients with Parkinson's disease," *Movement Disorders*, vol. 22, no. 16, p. S245, 2007.

[32] H. Zhao *et al.*, "Adaptive gait detection based on foot-mounted inertial sensors and multi-sensor fusion," *Information Fusion*, vol. 52, no. January, pp. 157–166, 2019, doi: 10.1016/j.inffus.2019.03.002.

[33] M. A. Case, H. A. Burwick, K. G. Volpp, and M. S. Patel, "Accuracy of smartphone applications and wearable devices for tracking physical activity data," *JAMA - Journal of the American Medical Association*, vol. 313, no. 6. 2015. doi: 10.1001/jama.2014.17841.

[34] J. M. Balto, D. L. Kinnett-Hopkins, and R. W. Motl, "Accuracy and precision of smartphone applications and commercially available motion sensors in multiple sclerosis," *Multiple Sclerosis Journal - Experimental, Translational and Clinical*, vol. 2, 2016, doi: 10.1177/2055217316634754.

[35] E. Chiauzzi, C. Rodarte, and P. DasMahapatra, "Patient-centered activity monitoring in the self-management of chronic health conditions," *BMC Medicine*, vol. 13, no. 1, 2015, doi: 10.1186/s12916-015-0319-2.

[36] D. J. Geerse, B. H. Coolen, and M. Roerdink, “Kinematic validation of a multi-Kinect v2 instrumented 10-meter walkway for quantitative gait assessments,” *PLoS ONE*, vol. 10, no. 10, Oct. 2015, doi: 10.1371/journal.pone.0139913.

[37] D. J. Geerse, M. Roerdink, J. Marinus, and J. J. van Hilten, “Assessing Walking Adaptability in Parkinson’s Disease: ‘The Interactive Walkway,’” *Frontiers in Neurology*, vol. 9, Dec. 2018, doi: 10.3389/fneur.2018.01096.

[38] M. C. Kirkland *et al.*, “Bipedal hopping timed to a metronome to detect impairments in anticipatory motor control in people with mild multiple sclerosis,” *Clinical Biomechanics*, vol. 55, pp. 45–52, Jun. 2018, doi: 10.1016/j.clinbiomech.2018.04.009.

[39] M. C. Kirkland *et al.*, “Bipedal Hopping Reveals Evidence of Advanced Neuromuscular Aging Among People With Mild Multiple Sclerosis,” *Journal of Motor Behavior*, vol. 49, no. 5, pp. 505–513, Sep. 2017, doi: 10.1080/00222895.2016.1241750.

[40] C. Leone *et al.*, “Effects of rehabilitation on gait pattern at usual and fast speeds depend on walking impairment level in multiple sclerosis,” *International Journal of MS Care*, vol. 20, no. 5, 2018, doi: 10.7224/1537-2073.2015-078.

[41] N. D. Chiaravalloti and J. DeLuca, “Cognitive impairment in multiple sclerosis,” *The Lancet Neurology*, vol. 7, no. 12, pp. 1139–1151, 2008, doi: 10.1016/S1474-4422(08)70259-X.

[42] L. Hemmett, J. Holmes, M. Barnes, and N. Russell, “What drives quality of life in multiple sclerosis?,” *QJM - Monthly Journal of the Association of Physicians*, vol. 97, no. 10, pp. 671–676, 2004, doi: 10.1093/qjmed/hch105.

[43] E. Buckley, C. Mazzà, and A. McNeill, “A systematic review of the gait characteristics associated with Cerebellar Ataxia,” *Gait and Posture*, vol. 60, no. August 2017, pp. 154–163, 2018, doi: 10.1016/j.gaitpost.2017.11.024.

[44] C. E. P. van Munster and B. M. J. Uitdehaag, “Outcome Measures in Clinical Trials for Multiple Sclerosis,” *CNS Drugs*, vol. 31, no. 3, pp. 217–236, 2017, doi: 10.1007/s40263-017-0412-5.

[45] A. Buoite Stella, M. E. Morelli, F. Giudici, A. Sartori, P. Manganotti, and P. E. di Prampero, “Comfortable walking speed and energy cost of locomotion in patients with multiple sclerosis,” *European Journal of Applied Physiology*. Springer, 2020. doi: 10.1007/s00421-019-04295-3.

[46] A. Middleton, G. D. Fulk, T. M. Herter, M. W. Beets, J. Donley, and S. L. Fritz, “Self-selected and maximal walking speeds provide greater insight into fall status than walking speed reserve among community-dwelling older adults,” *American Journal of Physical Medicine and Rehabilitation*, vol. 95, no. 7, pp. 475–482, 2016, doi: 10.1097/PHM.0000000000000488.

[47] A. Middleton, G. D. Fulk, M. W. Beets, T. M. Herter, and S. L. Fritz, “Self-selected walking speed in predictive of daily ambulatory activity in older adults,” *Journal of aging and physical activyti*, vol. 24, no. 2, pp. 214–222, 2015, doi: 10.1123/japa.2015-0104.Self-Selected.

[48] J. L. Preiningerova, K. Novotna, J. Rusz, L. Sucha, E. Ruzicka, and E. Havrdova, “Spatial and temporal characteristics of Gait as outcome measures in multiple sclerosis (EDSS 0 to 6.5),” *Journal of NeuroEngineering and Rehabilitation*, vol. 12, no. 1, pp. 1–7, 2015, doi: 10.1186/s12984-015-0001-0.

[49] D. A. Henning, E. M. Edwards, M. Ansara, and N. E. Fritz, “Validating the walking while talking test to measure motor, cognitive, and dual-task performance in ambulatory individuals with multiple sclerosis,” *Multiple Sclerosis and Related Disorders*, vol. 54, no. June, p. 103123, 2021, doi: 10.1016/j.msard.2021.103123.

[50] M. J. Hohol, E. J. Orav, and H. L. Weiner, “Disease Steps in multiple sclerosis: A simple approach to evaluate disease progression,” *Neurology*, vol. 45, no. 2, pp. 251–255, Feb. 1995, doi: 10.1212/WNL.45.2.251.

[51] M. C. Kirkland, E. M. Wallack, S. N. Rancourt, and M. Ploughman, “Comparing Three Dual-Task Methods and the Relationship to Physical and Cognitive Impairment in People with Multiple Sclerosis and Controls,” *Multiple Sclerosis International*, vol. 2015, pp. 1–7, 2015, doi: 10.1155/2015/650645.

[52] M. C. Kirkland, K. P. Wadden, and M. Ploughman, “Bipedal hopping as a new measure to detect subtle sensorimotor impairment in people with multiple sclerosis,” *Disability and Rehabilitation*, vol. 0, no. 0, pp. 1–12, 2020, doi: 10.1080/09638288.2020.1820585.

[53] W. S. Kim and E. Y. Kim, “Comparing self-selected speed walking of the elderly with self-selected slow, moderate, and fast speed walking of young adults,” *Annals of Rehabilitation Medicine*, vol. 38, no. 1, pp. 101–108, 2014, doi: 10.5535/arm.2014.38.1.101.

[54] C. Berlet, “Machine Learning: A Bayesian and Optimization Perspective-Introduction,” *Trumping Democracy*, no. 2015, pp. 3–23, 2020, doi: 10.4324/9781315438412-2.

[55] M. and D. S. J. Cunningham Pádraig and Cord, “Supervised Learning,” in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, P. Cord Matthieu and Cunningham, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 21–49. doi: 10.1007/978-3-540-75171-7_2.

[56] D. R. Cox, “The Regression Analysis of Binary Sequences,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[57] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[58] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[59] A. Li Stan Z. and Jain, Ed., “AdaBoost,” in *Encyclopedia of Biometrics*, Boston, MA: Springer US, 2009, p. 9. doi: 10.1007/978-0-387-73003-5_825.

[60] Z. Ghahramani, “Unsupervised Learning,” in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.

[61] R. S. Mcginnis *et al.*, “A machine learning approach for gait speed estimation using skin-mounted wearable sensors : From healthy controls to individuals with multiple sclerosis,” pp. 1–16, 2017.

[62] A. Supratak, G. Datta, A. R. Gafson, R. Nicholas, and Y. Guo, “Remote Monitoring in the Home Validates Clinical Gait Measures for Multiple Sclerosis,” vol. 9, no. July, pp. 1–9, 2018, doi: 10.3389/fneur.2018.00561.

[63] “Psychophysical bases of perceived exertion”.

[64] J. Neuroengineering *et al.*, “Inertial sensor - based gait parameters reflect patient - reported fatigue in multiple sclerosis,” *Journal of NeuroEngineering and Rehabilitation*, pp. 1–9, 2020, doi: 10.1186/s12984-020-00798-9.

[65] R. Kaur, Z. Chen, R. Motl, M. E. Hernandez, and R. Sowers, “Predicting Multiple Sclerosis from Gait Dynamics Using an Instrumented Treadmill – A Machine Learning Approach,” *IEEE Transactions on Biomedical Engineering*, 2020, doi: 10.1109/TBME.2020.3048142.

[66] A. R. Chaves *et al.*, “Asymmetry of Brain Excitability: A New Biomarker that Predicts Objective and Subjective Symptoms in Multiple Sclerosis,” *Behavioural Brain Research*, vol. 359, no. October 2018, pp. 281–291, 2019, doi: 10.1016/j.bbr.2018.11.005.

[67] D. A. Galloway *et al.*, “miR-223 promotes regenerative myeloid cell phenotype and function in the demyelinated central nervous system,” *Glia*, vol. 67, no. 5, pp. 857–869, 2019, doi: 10.1002/glia.23576.

[68] G. Severini *et al.*, “Evaluation of Clinical Gait Analysis parameters in patients affected by Multiple Sclerosis: Analysis of kinematics,” *Clinical Biomechanics*, vol. 45, no. November 2016, pp. 1–8, 2017, doi: 10.1016/j.clinbiomech.2017.04.001.

[69] T. Bushnik, “Expanded Disability Status Scale,” *Encyclopedia of Clinical Neuropsychology*, pp. 1363–1365, 2018, doi: 10.1007/978-3-319-57111-9_1805.

[70] G. A. Phillips *et al.*, "Responder definition of the Multiple Sclerosis Impact Scale physical impact subscale for patients with physical worsening," *Multiple Sclerosis Journal*, vol. 20, no. 13, pp. 1753–1760, 2014, doi: 10.1177/1352458514530489.

[71] "Multiple Sclerosis Impact Scale (MSIS-29)," 2000.

[72] "GAITRite Electronic Walkway Technical Reference (47DevM2)."

[73] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[74] A. S. Ranawat, M. A. Gaudiani, P. A. Slullitel, J. Satalich, and B. J. Rebolledo, "Foot Progression Angle Walking Test: A Dynamic Diagnostic Assessment for Femoroacetabular Impingement and Hip Instability," *Orthopaedic Journal of Sports Medicine*, vol. 5, no. 1, Jan. 2017, doi: 10.1177/2325967116679641.

[75] J. J. Sosnoff, B. M. Sandroff, and R. W. Motl, "Quantifying gait abnormalities in persons with multiple sclerosis with minimal disability," *Gait and Posture*, vol. 36, no. 1, pp. 154–156, May 2012, doi: 10.1016/j.gaitpost.2011.11.027.

[76] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Computer Science*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.

[77] G. Menardi and N. Torelli, *Training and assessing classification rules with imbalanced data*, vol. 28, no. 1. 2014. doi: 10.1007/s10618-012-0295-5.

[78] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed. Wiley-IEEE Press, 2013.

[79] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the International Joint Conference on Neural Networks*, no. 3, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.

[80] A. L. Hof, "Scaling gait data to body size.," *Gait and Posture*. pp. 222–223, 1996.

[81] M. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[82] M. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[83] E. R. Girden, "ANOVA: Repeated measures," *SAGE Publications, Inc*, 1992.

[84] Fabian Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 9, pp. 2825–2830, 2011, doi: 10.1289/EHP4713.

[85] K. M. Ting, "Confusion Matrix," 2017, [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_50

[86] S. M. Piryonesi, S. Rostampour, and S. A. Piryonesi, "Predicting falls and injuries in people with multiple sclerosis using machine learning algorithms," *Multiple Sclerosis*

and Related Disorders, vol. 49, no. December 2020, p. 102740, 2021, doi: 10.1016/j.msard.2021.102740.

[87] J. S. Brach, J. E. Berlin, J. M. Vanswearingen, A. B. Newman, and S. A. Studenski, "Too much or too little step width variability is associated with a fall history in older persons who walk at or near normal gait speed," 2005, doi: 10.1186/1743.

[88] G. L. Widener and D. D. Allen, "Measurement characteristics and clinical utility of the 29-item Multiple Sclerosis Impact Scale.," *Arch Phys Med Rehabil*, vol. 95, no. 3, pp. 593–594, 2014, doi: 10.1016/j.apmr.2013.07.008.