2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822219

Data Augmentation for Classifying Multiple Sclerosis Severity through Inertial Measurement Unit-Based Gait Analysis

Jessica B. Li
Department of Statistics
Department of Computer Science
Northwestern University
Evanston, USA
jessicali2025@u.northwestern.edu

John W. Farrell III
Department of Health and Human Performance
Texas State University
San Marcos, USA
jwf77@txstate.edu

Damian Valles
Ingram School of Engineering
Texas State University
San Marcos, USA
dvalles@txstate.edu

***Abstract*—Gait impairments are highly prevalent in persons with multiple sclerosis (PwMS), contributing to difficulties in daily activities. Disability is commonly assessed using the Expanded Disability Status Scale (EDSS). However, the EDSS lacks detailed gait analysis for assessing the severity of multiple sclerosis. To address this, gait analysis tools such as inertial measurement units are commonly used to understand walking patterns in PwMS. Another concern is that collecting sufficient gait data becomes challenging due to limited participation in studies. This research acknowledges these limitations and proposes using a variational autoencoder to address this issue. Additionally, the study explores the feasibility of classification models aimed at assisting the quantification of disability of PwMS based on individuals' gait patterns.**

***Index Terms*—multiple sclerosis, data augmentation, variational autoencoder, inertial measurement unit, machine learning**

I. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system resulting in demyelination of neurons. The resulting deterioration of the myelin sheath exposes the axon of the neuron, causing the inhibition of the propagation of action potentials from the central nervous system to the peripheral tissue. Remyelination can occur to the myelin sheath, though only to a small degree, but a lesion, or sclerosis, is often left [1]. Due to the disruption of the propagation of action potentials, persons with MS (PwMS) experience many symptoms, often influenced by the location of the lesion and the severity of the inflammatory reaction. Common symptoms reported by PwMS include impaired vision, balance, muscle control, and sensation [2].

Currently, the severity of MS is quantitatively scored using the Expanded Disability Status Scale (EDSS). EDSS is an accepted instrument to assess the severity and progression of MS and is regularly used to note the effectiveness of therapeutic interventions [3]. The EDSS scoring system scales the severity level for MS from a rating of 0.0 to 10.0 in increments of 0.5. A rating of 0.0 equates to the individual having normal functions with no disability, and a rating of 10.0 equates to the individual being deceased due to MS [4]. The overall rating is calculated through evaluations of optic, brainstem, pyramidal, cerebellar, sensory, bowel/bladder, cerebral functions, and walking capability. Each of these categories has its predetermined scoring criteria. While EDSS scores increase in increments of 0.5, the scale is not continuous but is influenced by differing factors. Specifically, walking ability significantly impacts scores above 3 [5]. From the scoring rubric, EDSS ratings under 0.0 to 3.5 note that the person is fully ambulatory without aid and does not specify a limit to the distance a person can walk. Scores of 4.0 to 5.5 describe the individual as fully ambulatory without aid. Individuals can walk without rest for distances of 500 meters for 4.0, 300 meters for 4.5, 200 meters for 5.0, and 100 meters for 5.5. EDSS ratings past 5.5 state that the individual requires assistance with mobility or is completely restricted to a bed or wheelchair [6].

The loss of balance and muscle control contribute to the gait patterns of PwMS, such that more than 50% of people with MS have balance and walking difficulties [7]. Additionally, gait is one of the first and most common impairments to occur in the early stages of MS [8] [9]. Therefore, it is important to assess the gait characteristics of PwMS to determine trends in gait abnormalities. However, the EDSS rating system offers a surface-level assessment of walking ability, such that the PwMS either reports or has its maximum unassisted walking distance (in meters) measured by the neurologist. Many PwMS who choose to self-report their walking distance underestimates the distance, resulting in an incorrect EDSS score [10]. Previous studies have utilized sensor systems to capture gait parameters for people with MS, in which there were noticeable changes to parameters including step length, double support time, and walking speed across increasing EDSS levels and compared to healthy control participants [11]. Regarding groups of PwMS with worsening EDSS scores, there is a significant association between changes to medio-lateral stride

This project was funded by the National Science Foundation - Research Experiences for Undergraduates (NSF-REU) under Grant #2150135

regularity and disease progression [12]. Additionally, wearable sensors are an accessible alternative to more expensive gait measurement instruments [13] [14].

Other studies have used machine learning algorithms to classify PwMS from healthy controls [15] [16]. However, these studies rely on instrumented walkways and treadmills to measure walking characteristics, and few studies utilize Inertial Measurement Units (IMUs) to extract data for machine learning classification of MS. Another problem that arises is the limitation in access to MS data due to the unavailability of original datasets, time-consuming nature of data collection, and monetary restrictions in obtaining data or accessing commercial datasets [17] [18]. These limitations are influenced by such factors of geography, where the prevalence of PwMS increases in areas further away from the equator, and accessibility, where individuals would need to consider transportation and costs of transportation to available clinics that specialize in MS treatment [19]. Even without these limitations, with researchers looking into using and combining external datasets, existing real-world evidence (RWE) data for MS does not follow a standardized data collection and processing protocol [20].

Thus, the purpose of the current investigation was to determine if data generation is a viable solution to limited data resources in healthcare and if machine learning methods and comprehensive gait evaluation can more accurately assess the severity of MS. It was hypothesized that using neural networks to generate data from existing health data is a reliable option when access to data is limited. Additionally, it was hypothesized that the classification of gait patterns using machine learning can act as a complementary assessment to current standard methods of the EDSS.

II. Data Collection and Preprocessing

A. Data Collection

A control group of 8 healthy individuals was recruited to participate in the current investigation[1]. Participants were asked to complete the 6-Minute Walking Test (6MWT) once with their usual gait and a second time with an impairment to simulate the gait patterns of a PwMS. During both assessments, participants wore inertial measurement units to collect biomechanical data for gait analysis.

The 6-Minute Walk Test [21] is a standard functional assessment for PwMS used to evaluate walking endurance. A distance of 30 meters was marked out on a flat surface for the assessment. Participants were instructed to walk at their own pace to the end of the marked-out course, perform a U-turn, and walk back to the starting point, completing as many loops as possible in 6 minutes. The participants completed the 6MWT two times, once with their normal gait pattern and second time with a gait pattern to mimic a person with MS. Previous studies that have examined gait patterns in PwMS have observed reduced motion in the hips, knee joints, and ankle joints [22]. Thus, to mimic similar restrictions in motion, our gait pattern prompts included several variations of keeping one or both knees locked while walking, such as prompting the participant to keep one leg straight while walking, walking normally but locking one knee once the foot contacts the ground, and keeping both legs straight while walking.

IMU sensors consist of accelerometers, gyroscopes, and magnetometers measuring acceleration, rotational motion, and orientation across a three-dimensional plane [23]. Following previous studies, the sensors were placed on the lower extremities as shown in Figure 1 to assess the gait characteristics of the participants [24]. We utilize four Noraxon Myomotion IMU sensors strapped to the left and right foot and the left and right shank to measure the gait patterns of the participants. These four sensors captured 112 columns of data regarding foot pitch, foot roll, acceleration, orientation, trajectory, and contact. The Noraxon Myomotion sensors recorded these features with a frequency of 100 Hz. We utilize nine columns of the dataset that indicate the foot position and ground contact over the allotted time. Only the first and last minutes of the 6MWT were recorded to reduce the amount of data collected for each person. Sensors were calibrated according to the manufacturer's recommendations before each iteration of the 6MWT.

Fig. 1. Noraxon Myomotion IMU and Placement on Lower Body

B. Preprocessing

Each iteration of the 6MWT was then exported to a single CSV file. The data was processed into identifiable gait parameters using the given columns from the raw dataset:

- 'time'
- 'Segments-Foot RT-Contact'
- 'Segments-Foot LT-Contact'
- 'Trajectories-Heel back RT-x (mm)'
- 'Trajectories-Heel back LT-x (mm)'
- 'Trajectories-Heel back RT-y (mm)'
- 'Trajectories-Heel back LT-y (mm)'
- 'Trajectories-Heel back RT-z (mm)'
- 'Trajectories-Heel back LT-z (mm)'

The gait parameters [13] extracted from the raw data and their definitions are as follows:

- Stride length: distance between a foot's heel position at initial contact and its position at subsequent contact
- Step length: distance between the heel positions of opposing feet when each is on the ground (left step length

[1]The research protocol was reviewed and approved by the Institutional Review Board (IRB) at Texas State University under approval number #8289. Written informed consent was obtained from all individual participants included in the study.

is when the left foot is forward; right step length is when the right foot is forward)
- Step duration: time interval between one foot's initial ground contact and the opposite foot's ground contact
- Cadence: the number of steps taken over a minute
- Single support interval: time interval for when one foot is on the ground
- Double support interval: time interval for when both feet are on the ground

First, we converted the three-dimensional coordinates from millimeters to meters and then used the columns to calculate the left and right foot gait parameters. These calculations were split into the first and last minute and averaged to create a dataset of 22 continuous variables. This process was completed by implementing the algorithm in Figure 2.

Algorithm 1: Gait Parameter Calculation

```
Set previous heel positions to first entry
for each entry in dataset do
    Set left and right heel positions to entry
    if left heel strike occurs then
        Calculate step length with left heel position and previous right heel position
        Calculate step duration
        Calculate stride length with left heel position and previous left heel position
        Increment cadence
        Record step length, step duration, and stride length
        Update previous left heel position to current left heel position
    if right heel strike occurs then
        Calculate step length with right heel position and previous left heel position
        Calculate step duration
        Calculate stride length with right heel position and previous right heel position
        Increment cadence
        Record step length, step duration, and stride length
        Update previous right heel position to current right heel position
    if both feet are not on the ground then
        Record double support interval if time started
    else
        Start time
    if left foot is on the ground then
        Record left single support interval if time started
    else
        Start left foot time
    if right foot is on the ground then
        Record right single support interval if time started
    else
        Start right foot time
Calculate averages for the first 60 seconds
Calculate averages for the last 60 seconds
```

Fig. 2. Gait Parameter Pseudocode

Each walking impairment prompt was assessed and assigned a score based on the Expanded Disability Status Scale (EDSS) by a certified EDSS rater. The impairments reflected EDSS scores of 3.5 and 4.0, so data reflecting scores of 4.5, 5.0, and 5.5 were generated from the original data. The calculated values for the last minute were set to 0, as EDSS scores of 4.5 or higher indicate an inability to walk 500 meters without rest, while healthy individuals can walk 400 to 700 meters in six minutes [25].

The final prepared dataset consisted of 22 continuous variables of the gait parameters and one categorical variable of the associated EDSS score. Length calculations were rounded to the nearest thousandth, time calculations were rounded to the nearest hundredth, and cadence was recorded in integers.

III. Methodology

A. Data Generation

To resolve the problem of limited access to participant data, we chose to use data augmentation. For data augmentation, a generative artificial intelligence model is used to create synthetic data to train our classification model. We implemented a variational autoencoder (VAE) due to its ability to learn and define a latent space through the distribution of the original data, such that randomly generated noise can then be decoded to generate new data. VAEs have been widely utilized in previous studies for generating data, including text modeling, molecular structures, handwritten digits, and images of faces [26].

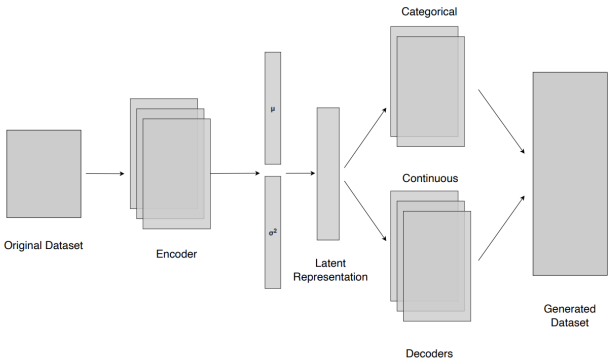

Fig. 3. Variational Autoencoder Architecture

The prepared dataset was first separated into categorical and continuous variables, where continuous features were transformed to follow a uniform distribution using quantile information and the categorical feature was converted with one-hot encoding. The architecture of the VAE is shown in Figure 3. The components of the VAE are split into two parts: the encoder and the decoder. Our encoder comprises an input layer, a hidden layer utilizing the Rectified Linear Unit (ReLU) activation function, and an output layer. We implement two decoders: one for categorical data and one for continuous data. The categorical decoder consists of an input layer followed by an output layer with softmax activation. The continuous decoder comprises an input layer, a hidden layer employing the ReLU activation function, and an output layer. During training, we use Adam with a learning rate of 0.0001 to minimize the given loss function:

$$\begin{aligned} \text{Loss} = &\int q(z|x) \log \frac{q(z|x)}{p(z)} dx \\ &+ \frac{1}{n} \sum_{i=1}^{n} (x_{\text{cont},i} - \hat{x}_{\text{cont},i})^2 \\ &- \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{\text{cat},i,j} \log \left(p_{\text{cat},i,j}\right) \end{aligned} \tag{1}$$

where $q(z|x)$ is the approximated posterior distribution, $p(z)$ is the prior distribution, n is the number of samples in the dataset, k is the number of classes for the categorical variable,

$x_{\text{cont},i}$ is the true value for the i^{th} continuous variable, $\hat{x}_{\text{cont},i}$ is the predicted value for the i^{th} continuous variable, $x_{\text{cat},i,j}$ is the indicator function for the i^{th} variable in the j^{th} class, and $p_{\text{cat},i,j}$ is the predicted probability for the i^{th} variable to belong to the j^{th} class. The first term of the loss function is the KL divergence term, which measures how closely the distribution of the approximated posterior, $q(z|x)$, matches the distribution of the prior, $p(z)$. Second, we add two reconstruction loss terms, with the first being the mean squared error for the continuous variables and the second being the categorical cross entropy for the categorical variables.

After training the model, data is generated by sampling latent variables from a standard Gaussian distribution. The latent variables were then decoded and inversely transformed to their original representation, resulting in a new dataset of generated samples.

Ultimately, we intend for the generated data to represent the original dataset fairly. Specifically, we used the two-sided Kolmogorov-Smirnov test for two samples to compare distributions for the original and generated data for each continuous variable. The formula for the KS test is provided below in Equation 2:

$$D_{mn} = \left(\frac{mn}{m+n}\right)^{1/2} \sup_x |F_m(x) - G_n(x)| \qquad (2)$$

where $F(x)$ is the empirical distribution function for the original dataset, $G(x)$ is the empirical distribution function for the generated dataset, m is the sample size of the original dataset, n is the sample size of the generated dataset, and D is the test statistic with mn degrees of freedom. The two-sample Kolmogorov-Smirnov statistic tests the null hypothesis that two empirical distribution functions are equal instead of the alternative that the two empirical distribution functions are not the same. We state our null hypothesis and alternative hypothesis to be:

$$H_0 : F(x) = G(x) \text{ vs. } H_1 : F(x) \neq G(x) \qquad (3)$$

We set our significance level to 0.05, which states that the two datasets were not drawn from the same distribution if we get a p-value of less than 0.05.

B. Classification Models and Implementation

Our model aims to predict the EDSS score given to a PwMS based on the decided gait features. We implemented multiple models in consideration of the nature of our data and the suitability for future use in a clinical setting. These classification models included Random Forest, Gaussian Naive Bayes, Logistic Regression, and Gradient Boosting.

Before implementing the four models, features were standardized to follow a normal distribution with 0 mean and unit variance before training and prediction occurred. The 500-sample dataset was split into 64-16-20 training, validation, and testing splits. The hyperparameters of the four classification models were tuned based on the evaluation metrics of the validation set.

C. Evaluation

The performance of the classification models was evaluated by calculating the respective balanced accuracy, precision, recall, and f1-score for each class. We utilized confusion matrices to visualize each model's prediction accuracy. Then, we tested the model on the original dataset from the best-performing classification model to evaluate the reliability of training the model on generated data.

IV. Results

A. Data Generation

Figure 4 shows 23 histograms visualizing data distribution across the 23 variables to compare how closely the generated dataset represents the original dataset. Each graph represents one of the 23 variables, in which the x-axis displays the range of values for the given variable, and the y-axis displays the density for that value in the dataset. The bars shaded orange show the distribution for the generated dataset, and the bars shaded blue show the distribution for the original dataset.

TABLE I
KS Test Statistic and P-value for Comparing Original and Generated Datasets

Gait Parameter	D_{mn}	P-value
LF M1 Stride Length	0.210	0.257
LF M6 Stride Length	0.248	0.146
RF M1 Stride Length	0.222	0.189
RF M6 Stride Length	0.240	0.168
LF M1 Step Length	0.214	0.225
LF M6 Step Length	0.284	0.065
RF M1 Step Length	0.314	**0.026**
RF M6 Step Length	0.196	0.346
LF M1 Step Duration	0.270	0.052
LF M6 Step Duration	0.110	0.853
RF M1 Step Duration	0.250	0.050
RF M6 Step Duration	0.126	0.714
M1 DSI	0.148	0.388
M6 DSI	0.120	0.674
LF M1 SSI	0.250	0.097
LF M6 SSI	0.150	0.527
RF M1 SSI	0.302	**0.018**
RF M6 SSI	0.130	0.719
LF M1 Cadence	0.296	**0.009**
LF M6 Cadence	0.236	0.129
RF M1 Cadence	0.290	**0.023**
RF M6 Cadence	0.228	0.160

Abbreviations: LF-left foot; RF-right foot; M1-first minute; M6-last minute; DSI-double support interval; SSI-single support interval

Table I is used to quantitatively compare the two datasets using the KS test statistic and associated p-values. From the table, the variables 'RF M1 Step Length', 'RF M1 SSI', 'LF M1 Cadence', and 'RF M1 Cadence' all correspond to p-values lower than 0.05 while the other 18 gait parameters have p-values above 0.05. This means there is a significant difference in distributions between the original and generated

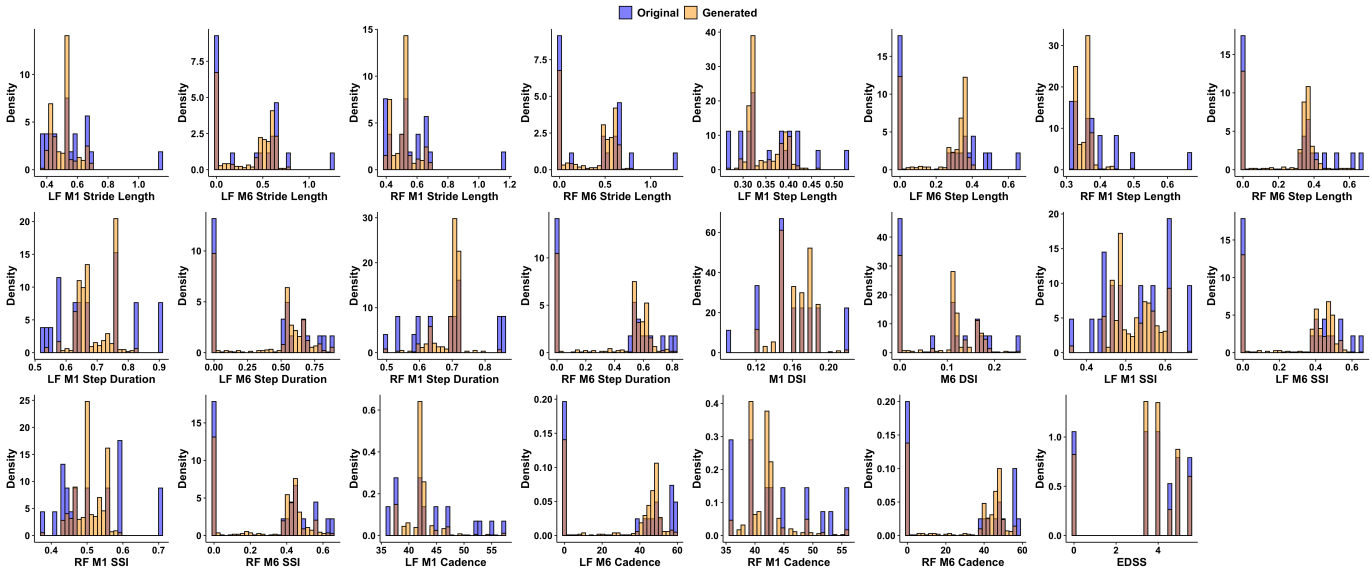

Fig. 4. Histograms Comparing Original and Generated Datasets Across the 23 Variables

Abbreviations: LF-left foot; RF-right foot; M1-first minute; M6-last minute; DSI-double support interval; SSI-single support; EDSS-Expanded Disability Status Scale

datasets for these four gait parameters. Since the remaining gait parameters have p-values above 0.05, there is no sufficient evidence to reject the null hypothesis and state that the original and generated data distributions differ.

We compare these obtained p-values to the histograms in Figure 4 to assess these differences visually. It is evident that although the generated dataset follows the overall distributions across the 23 variables, the VAE tended to generate more data clustering around the means of the variables. Thus, in the cases of a first-minute right foot step length, single support interval, cadence, and left foot cadence, the VAE did not fully capture the variability from the original dataset.

Another aim for the VAE was to correlate scores of 4.5, 5.0, and 5.5 with being unable to walk 500 meters without rest or sustain a 6-minute walk. In the generated dataset, this is observed with zero values for the gait measurements recorded at the last minute. Figure 5 shows the occurrence of zeros and other values for the eleven gait measurements in EDSS scores of 4.5 and above. The histograms reveal that zero values are most common for all gait parameters measured at the last minute of the 6MWT, with fewer frequent values above zero.

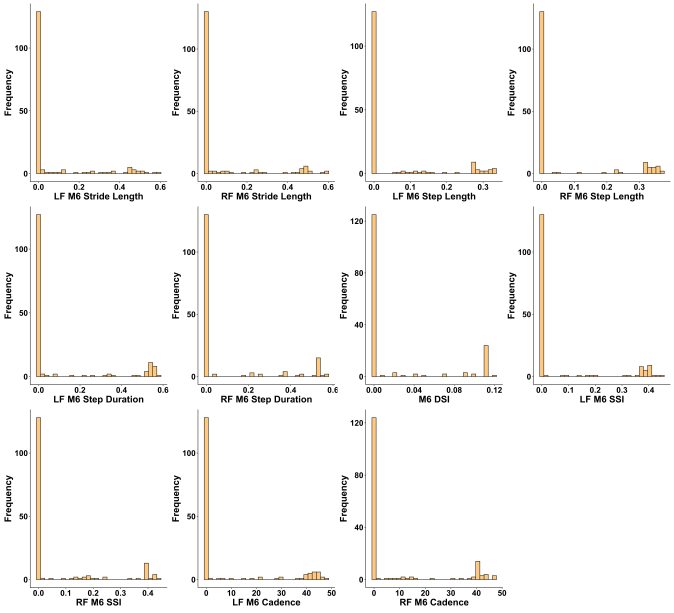

Fig. 5. Generated Measurements for Scores 4.5 and Above in the Last Minute

Abbreviations: LF-left foot; RF-right foot; M6-last minute; DSI-double support interval; SSI-single support

B. Model Performance

Based on the confusion matrices in Figure 6, the Random Forest classifier correctly predicted 85 ratings, the Logistic Regression classifier correctly predicted 86 ratings, the Gaussian Naive Bayes classifier correctly predicted 80 ratings, and the Gradient Boosting classifier correctly predicted 85 ratings out of a total of 100 samples. All four models exhibit similar patterns in prediction errors in which higher true ratings of 5.0 and 5.5 were misclassified as ratings of 3.5.

Precision, recall, and F1-scores were calculated for each class in Table II, with most scores exceeding 0.70. Notably, the Gradient Boosting model had one score below 0.70, while the Gaussian Naive Bayes classifier had two scores below 0.70. In Table III, the Gaussian Naive Bayes classifier had the lowest balanced accuracy score of 0.761 among the four models. The other three models performed comparably, with balanced accuracy scores around 0.85. The precision, recall, and F1-scores for the Random Forest, Logistic Regression, and

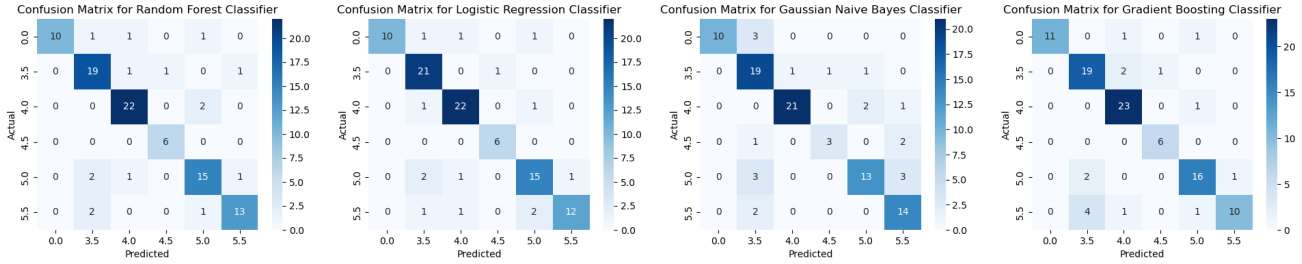

Fig. 6. Test Set Confusion Matrices

TABLE II
PERFORMANCE METRICS FOR TEST SET

Model	EDSS Rating	Precision	Recall	F1-Score
Random Forest	0.0	1	0.77	0.87
	3.5	0.79	0.86	0.83
	4.0	0.88	0.92	0.90
	4.5	0.86	1	0.92
	5.0	0.79	0.79	0.79
	5.5	0.87	0.81	0.84
Logistic Regression	0.0	1	0.77	0.87
	3.5	0.81	0.95	0.88
	4.0	0.88	0.92	0.90
	4.5	0.86	1	0.92
	5.0	0.79	0.79	0.79
	5.5	0.92	0.75	0.83
Gaussian Naive Bayes	0.0	1	0.77	0.87
	3.5	0.68	0.86	0.76
	4.0	0.95	0.88	0.91
	4.5	0.75	0.5	0.6
	5.0	0.81	0.68	0.74
	5.5	0.7	0.88	0.78
Gradient Boosting	0.0	1	0.85	0.92
	3.5	0.76	0.86	0.81
	4.0	0.85	0.96	0.9
	4.5	0.86	1	0.92
	5.0	0.84	0.84	0.84
	5.5	0.91	0.63	0.74

Gradient Boosting models ranged from 0.70 to 0.92. Generally, these three classification models demonstrated higher precision, recall, and F1-scores for EDSS ratings of 4.0 and 4.5, with the logistic regression model performing the best out of the four classifiers.

After testing the trained logistic regression model to predict EDSS scores from the original gait data, we achieved the metrics in Table IV. The model performs relatively well with a balanced accuracy of 0.74 and performs the best when

TABLE III
BALANCED ACCURACY FOR TEST SET

Model	Balanced Accuracy
Random Forest	0.859
Logistic Regression	0.863
Gaussian Naive Bayes	0.761
Gradient Boosting	0.856

TABLE IV
PERFORMANCE METRICS FOR LOGISTIC REGRESSION IN PREDICTING ORIGINAL DATASET EDSS SCORES

EDSS Rating	Precision	Recall	F1-Score	Balanced Accuracy
0.0	1	0.5	0.67	0.74
3.5	0.6	0.75	0.67	
4.0	0.8	1	0.89	
4.5	1	0.5	0.67	
5.0	0.6	1	0.75	
5.5	1	0.67	0.8	

predicting scores of 4.0, 5.0, and 5.5. We also provide a visual representation of the predictions with the confusion matrix in Figure 7, which shows that prediction errors made by the classifier fall within 0.5 of the true EDSS score.

V. DISCUSSION

To evaluate the efficacy of generating a larger dataset from an original, manually collected dataset, we utilize the Kolmogorov-Smirnov test for two samples and visualize the distributions with overlaid histograms. We then applied four machine learning classification algorithms to assess the practicality of using gait parameters and machine learning to assign an EDSS rating to PwMS. The results were promising, with most generated gait parameters matching the original data's distributions and three models achieving balanced accuracy scores above 80%.

We obtained p-values of 0.05 or greater for 18 gait parameters and p-values less than 0.05 for four gait parameters. The KS test was not employed to compare the distribution of EDSS scores as the EDSS scores do not follow a continuous distribution. The four gait parameters associated with p-values smaller

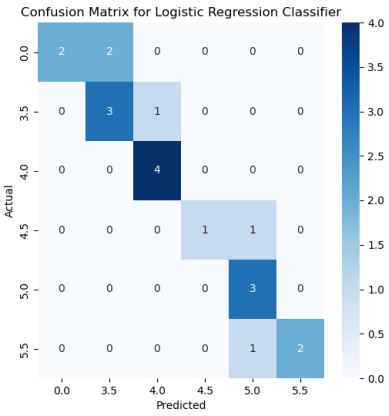

Fig. 7. Logistic Regression Confusion Matrix - Original Dataset

than 0.05 included the first-minute data for right foot step length, single support interval, cadence, and left foot cadence. A previous study determined walking cadence to vary depending on the height of the participant [27]. From the first-minute cadence histograms in Figure 4, the original dataset follows a slightly uniform distribution, which the participants' height variation can explain. The variability of walking cadences and the small size of the original dataset could contribute to the small p-value. Thus, further studies should consider standardizing cadence in accordance with height. We also focused on the VAE's ability to learn effectively and correlate scores of 4.5, 5.0, and 5.5 with the inability to walk 500 meters without rest. By accurately reflecting the gait characteristics of these disability levels, the model becomes more applicable to real gait data. Ultimately, our model holds promising results by capturing the underlying distributions for 18 of the 22 gait parameters. Considering these 18 parameters, the model performed well in reconstructing data with less variability, as the distribution is easier to represent and reconstruct.

We produced four working classification models from the generated dataset, namely using random forest, logistic regression, gaussian naive bayes, and gradient boosting algorithms, to predict the EDSS score from 22 gait characteristics. The logistic regression model performs best when we evaluate the four models on the precision, recall, and F1-scores between the six classes. This simplifies interpretation in a clinical setting, as logistic regression weights gait characteristics differently based on their significance in predicting an EDSS score. We also use a simpler dataset of 22 features so the logistic regression model can better generalize the given data. When testing the logistic regression classifier on the original dataset, we see a decrease in overall balanced accuracy, precision, recall, and f1-scores. However, the prediction errors fall closer to the true EDSS scores than the prediction results using the generated dataset, possibly due to similar gait characteristics between adjacent scores. This shows that our model can classify real data and complement other clinical assessment methods to improve the overall diagnosis.

From a practical standpoint, researchers with limited access to health-related datasets can use generative models for data augmentation since this alternative also protects personal health information. VAEs are a viable choice in the pool of generative models due to their ability to learn the data structure and prevalence for data or image generation tasks. Using a classification model to evaluate gait characteristics can provide a more comprehensive diagnosis for the disability, along with consideration of the current EDSS evaluation standards. The proposed logistic regression model is especially insightful for practice because clinicians can view how different features are weighted to influence the classification. As an attribute of being a probabilistic model, logistic regression also provides the probability of the given features being in the predicted class. This approach helps to mitigate the current black box problems in using artificial intelligence for medical fields.

VI. Future Work

Although we find our data augmentation and machine learning methods to be sufficient, we acknowledge there are limitations to the current investigation. The discrepancies from all four gait parameters can be attributed to our implemented VAE suffering from posterior collapse. The posterior collapse phenomenon is when the latent variables cannot capture sufficient information on the data. Our VAE model fails to represent the first-minute cadence, right foot step length, and single support interval parameters through the latent variables. Therefore, the current proposed model can be improved upon by experimenting with using solutions to prevent posterior collapse, such as a beta-VAE or Latent-Identifiable VAE (LIDVAE) model [28] [29]. Due to time constraints, we could not implement and test the performances of beta-VAE and LIDVAE models. Future studies could explore variations of the traditional VAE architecture to prevent posterior collapse and evaluate whether the generated data accurately represents all gait parameters from the original dataset. Additionally, collecting more data from PwMS would help the encoder better learn and represent the variation in gait characteristics. The performance of the classification models can be improved with more data. Recruiting PwMS with EDSS scores of 5.0 and 5.5 would help the model accurately discern walking characteristics, as these individuals are expected to exhibit more distinct walking patterns due to their shorter walking distances before needing to rest. Further research should consider a larger set of gait parameters to represent the gait cycle better. For example, the classification model can evaluate factors including ankle rotation and knee and hip range of motion.

VII. Conclusion

This research study introduces a solution for handling small datasets for health-related datasets. The proposed VAE architecture is demonstrated to be effective in data augmentation. Generally, the architecture of VAEs can be adapted to align with the characteristics of the original dataset. In our study, a shallow VAE with two separate decoders effectively

transformed Gaussian random samples into a larger dataset resembling the original collected data. From the generated data, we analyzed the performance of four classification models in distinguishing EDSS scores from relevant gait characteristics. The balanced accuracy and f1-scores of the models revealed promising results for utilizing machine learning methods to aid clinicians in rating the disability. Additionally, the model accurately classified the EDSS score or predicted within 0.5 of the actual score on the original dataset, demonstrating the feasibility of training the classification model on augmented data. Along with using IMUs, this process allows for a comprehensive and accessible assessment of walking disability for PwMS. Thus, the integration of IMUs and artificial intelligence in a clinical setting can provide an additional tool for improving the diagnosis of MS. Therefore, our initial results warrant further research in combining other modes of diagnosis for MS, including MRI scans, arm and leg strength, and vision tests to create a complete assessment of the disorder.

REFERENCES

[1] D. Tafti, M. Ehsan, and K. L. Xixis, "Multiple Sclerosis," in *StatPearls [Internet]*, Treasure Island, FL: StatPearls Publishing, Jan. 2024.

[2] "Multiple Sclerosis | National Institute of Neurological Disorders and Stroke."

[3] S. Meyer-Moock, Y.-S. Feng, M. Maeurer, F.-W. Dippel, and T. Kohlmann, "Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis," *BMC Neurology*, vol. 14, p. 58, Dec. 2014.

[4] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, pp. 1444–1444, Nov. 1983.

[5] J. Lizrova 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, p. 14, 2015.

[6] M. S. Trust, "Expanded Disability Status Scale (EDSS) | MS Trust."

[7] M. H. Cameron and Y. Nilsagard, "Balance, gait, and falls in multiple sclerosis," in *Handbook of Clinical Neurology*, vol. 159, pp. 237–250, Elsevier, 2018.

[8] C. L. Martin, B. A. Phillips, T. J. Kilpatrick, H. Butzkueven, N. Tubridy, E. McDonald, and M. P. Galea, "Gait and balance impairment in early multiple sclerosis in the absence of clinical disability," *Multiple Sclerosis Journal*, vol. 12, pp. 620–628, Sept. 2006.

[9] N. G. LaRocca, "Impact of Walking Impairment in Multiple Sclerosis: Perspectives of Patients and Care Partners," *The Patient: Patient-Centered Outcomes Research*, vol. 4, pp. 189–201, Sept. 2011.

[10] A. G. Skjerbæk, F. Boesen, T. Petersen, P. V. Rasmussen, E. Stenager, M. Nørgaard, P. Feys, M. L. Kjeldgaard-Jørgensen, L. G. Hvid, and U. Dalgas, "Can we trust self-reported walking distance when determining EDSS scores in patients with multiple sclerosis? The Danish MS hospitals rehabilitation study," *Multiple Sclerosis Journal*, vol. 25, pp. 1653–1660, Oct. 2019.

[11] L. Angelini, W. Hodgkinson, C. Smith, J. M. Dodd, B. Sharrack, C. Mazzà, and D. Paling, "Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting," *Journal of Neurology*, vol. 267, pp. 2897–2909, Oct. 2020.

[12] E. Gervasoni, D. Anastasi, R. Di Giovanni, C. Solaro, M. Rovaris, G. Brichetto, P. Confalonieri, A. Tacchino, I. Carpinella, and D. Cattaneo, "Uncovering Subtle Gait Deterioration in People with Early-Stage Multiple Sclerosis Using Inertial Sensors: A 2-Year Multicenter Longitudinal Study," *Sensors*, vol. 23, p. 9249, Nov. 2023.

[13] M. A. Laribi and S. Zeghloul, "Human lower limb operation tracking via motion capture systems," in *Design and Operation of Human Locomotion Systems*, pp. 83–107, Elsevier, 2020.

[14] M. Pau, S. Caggiari, A. Mura, F. Corona, B. Leban, G. Coghe, L. Lorefice, M. G. Marrosu, and E. Cocco, "Clinical assessment of gait in individuals with multiple sclerosis using wearable inertial sensors: Comparison with patient-based measure," *Multiple Sclerosis and Related Disorders*, vol. 10, pp. 187–191, Nov. 2016.

[15] K. Trentzsch, P. Schumann, G. Śliwiński, P. Bartscht, R. Haase, D. Schriefer, A. Zink, A. Heinke, T. Jochim, H. Malberg, and T. Ziemssen, "Using Machine Learning Algorithms for Identifying Gait Parameters Suitable to Evaluate Subtle Changes in Gait in People with Multiple Sclerosis," *Brain Sciences*, vol. 11, p. 1049, Aug. 2021.

[16] W. Hu, O. Combden, X. Jiang, S. Buragadda, C. J. Newell, M. C. Williams, A. L. Critch, and M. Ploughman, "Machine learning classification of multiple sclerosis patients based on raw data from an instrumented walkway," *BioMedical Engineering OnLine*, vol. 21, p. 21, Dec. 2022.

[17] C. S. Kwok, E.-A. Muntean, C. D. Mallen, and J. A. Borovac, "Data Collection Theory in Healthcare Research: The Minimum Dataset in Quantitative Studies," *Clinics and Practice*, vol. 12, pp. 832–844, Oct. 2022.

[18] I. R. I. Alberto, N. R. I. Alberto, A. K. Ghosh, B. Jain, S. Jayakumar, N. Martinez-Martin, N. McCague, D. Moukheiber, L. Moukheiber, M. Moukheiber, S. Moukheiber, A. Yaghy, A. Zhang, and L. A. Celi, "The impact of commercial health datasets on medical research and health-care algorithms," *The Lancet Digital Health*, vol. 5, pp. e288–e294, May 2023.

[19] G. C. Ebers and A. D. Sadovnick, "The Geographic Distribution of Multiple Sclerosis: A Review," *Neuroepidemiology*, vol. 12, no. 1, pp. 1–5, 1993.

[20] T. Ziemssen, D. Rothenbacher, J. Kuhle, and T. Berger, "Real-world-Evidenz: Vorteile und Limitationen am Beispiel der Multiplen Sklerose," *Der Nervenarzt*, vol. 88, pp. 1153–1158, Oct. 2017.

[21] "ATS Statement: Guidelines for the Six-Minute Walk Test," *American Journal of Respiratory and Critical Care Medicine*, vol. 166, pp. 111–117, July 2002.

[22] M. Coca-Tapia, A. Cuesta-Gómez, F. Molina-Rueda, and M. Carratalá-Tejada, "Gait Pattern in People with Multiple Sclerosis: A Systematic Review," *Diagnostics (Basel, Switzerland)*, vol. 11, p. 584, Mar. 2021.

[23] R. Kala, "An introduction to robotics," in *Autonomous Mobile Robots*, pp. 1–48, Elsevier, 2024.

[24] A. Vienne-Jumeau, F. Quijoux, P.-P. Vidal, and D. Ricard, "Wearable inertial sensors provide reliable biomarkers of disease severity in multiple sclerosis: A systematic review and meta-analysis," *Annals of Physical and Rehabilitation Medicine*, vol. 63, pp. 138–147, Mar. 2020.

[25] A. Chetta, A. Zanini, G. Pisi, M. Aiello, P. Tzani, M. Neri, and D. Olivieri, "Reference values for the 6-min walk test in healthy subjects 20–50 years old," *Respiratory Medicine*, vol. 100, pp. 1573–1578, Sept. 2006.

[26] J. M. Davila Delgado and L. Oyedele, "Deep learning with small datasets: using autoencoders to address limited datasets in construction management," *Applied Soft Computing*, vol. 112, p. 107836, Nov. 2021.

[27] D. A. Rowe, G. J. Welk, D. P. Heil, M. T. Mahar, C. D. Kemble, M. A. Calabró, and K. Camenisch, "Stride Rate Recommendations for Moderate-Intensity Walking," *Medicine & Science in Sports & Exercise*, vol. 43, pp. 312–318, Feb. 2011.

[28] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," Apr. 2018. arXiv:1804.03599 [cs, stat].

[29] Y. Wang, D. M. Blei, and J. P. Cunningham, "Posterior Collapse and Latent Variable Non-identifiability," Jan. 2023. arXiv:2301.00537 [cs, stat].