Wearables for movement state classification and Parkinson's Disease detection

Sarah Binder
CISC5800 Machine Learning
Fordham University
sbinder3@fordham.edu

***Abstract*— The use of inertial measurement unit (IMU) wearable devices poses an opportunity for robust health monitoring. Health monitoring encompasses a wide range of use cases and requires precision and specificity in deriving explainable actions from sensor data. One specific clinical use case of IMU wearable devices is symptom detection and monitoring in the context of Parkinson's Disease, a condition characterized by motor impairment. The current study serves two main aims. First, to validate the use of machine learning models in deriving specific and unique movement states. Second, to accurately predict Parkinson's Disease status from IMU data captured across various sensor locations. The first aim employs the use of Lasso Logistic Regression (Accuracy 0.9475, AUC 0.9955) and Support Vector Machine models (Accuracy 0.9511). The second aim is achieved (Accuracy 0.90, AUC 0.954) using a convolutional neural network (CNN).**

Keywords*— *inertial measurement unit (IMU), wearables, Parkinson's Disease, classification, deep-learning

I. Introduction

Inertial measurement unit (IMU) wearable devices are gaining recognition for their large potential to improve health and disease monitoring and detection. Parkinson's Disease (PD) is a neurodegenerative disorder characterized by gait impairment and motor abnormalities [1] and thus sensorized devices that detect motion patterns contain enormous potential for identifying Parkinsonian symptoms. Recently, there has been a large research interest in developing models to serve this purpose. The Parkinson's Disease Digital Biomarker DREAM Challenge crowdsourced approaches to identify the presence of PD and predict symptom severity from both a smartphone-based study and wrist accelerometer study, reporting an Area-under-the-Curve (AUC) of 0.87 for the model best at predicting PD status [2]. Another team tackled a similar problem with wrist-accelerometer data, reporting an AUC of 0.78 in distinguishing diagnosed PD from matched unaffected controls [3]. In combination with the clinical potential, wearable devices have become ubiquitous in the consumer health technology market, indicating a data source with untapped potential for continuous health monitoring and early-disease detection [4]. Given the promise of wearable IMU devices, the current project sought to accomplish two aims. First, to classify movement states of the Timed Up and Go (TUG) task, a task that includes hyper-specific motion states such as turning to sit in a chair, in order to further validate the utility of IMU data in characterizing movement states, even specific and complicated motions. This aim contributes to the literature surrounding the interpretability of wearable devices. The second aim is to accurately predict the prescence of Parkinson's Disease solely using IMU data.

II. Background

A. Data Description

1) The data used in the current research project was taken from the open-access Wearables for gait in Parkinson's Disease and age-matched controls (WearGait-PD) dataset, accessible via Synapse.org. The Wear-Gait PD study was a joint effort across four medical research teams, in which 61 total PD participants and 65 total age-matched controls performed standardized movement tasks. All participants completed ten motion tasks after being fit with IMU sensors in 13 locations on the body. The tasks took place on a sensorized walkway and were video recorded. All sources of data collection were time-synchronized with the sensorized walkway. The IMU sensors transmitted output at 100 Hz, translating to 0.01 second intervals. Video annotations labeled the movement state for each of these intervals [5]. The current project focused on the Timed Up and Go (TUG) task, in which participants began seated, stood up and walked to the end of the mat, turned at a marked location on the mat, walked back to the chair and sat down. The TUG task thus encompassed five unique movement states: sit-to-stand, sitting, turning, turn-to-sit, and walking. The first aim of the current project only utilized the TUG task, chosen for its unique, transitional motion states which would better test the true ability of a model to predict complex movements. Although other tasks involved turning, none other than the TUG task involved turn-to-sit, or sit-to-stand, so data augmentation with other tasks would only worsen the class imbalance, which is innate to the data given that these transitional states encompass brief and infrequent movements. For the second aim of the current project, data from three additional tasks were included to bolster the size of the training and testing sets. The selected tasks were the self-paced and hurried-pace walking tasks as well as the self-paced turn task which requires participants to turn 4 times, making both left and right turns.

The IMU sensor data included raw acceleration, free (linear) acceleration, angular velocity, magnetic field strength, velocity increment, and orientation change all in the X, Y and Z directions, in addition to roll, pitch, and yaw. For the purposes of narrowing the feature space and removing redundancy, the following raw features were selected for use in the current study, along all axes (X,Y,Z) when applicable: free acceleration, angular velocity, velocity increment, roll, pitch, and yaw.

III. Methodology

A. Preprocessing

For both aims, preprocessing steps included a standard pipeline of converting data types to numerical values, handling missing data by first dropping columns which contained majority missing data and then dropping remaining rows with missing data. All models employed standard scaling to normalize the data and included the entire available cohort (Parkinson's Disease patients and control subjects who passed the missing-data threshold) in both the testing and training stages to ensure generalizability. Every model was executed using GroupK-fold cross validation to ensure that windowed data from the same participant was not split between training and testing sets, which would cause data leakage and unreliable results.

B. Aim 1 - Feature Engineering and Model Development

To address the first aim, 1-second sliding windows with 50% overlap were computed, resulting in roughly 100 windows of data per participant. Features were engineered by computing summary statistics of the windowed data. This included mean, standard deviation, minimum, maximum, root mean squared (RMS) chosen to provide a measure of magnitude, in addition to peak counts to capture oscillatory patterns, and zero-crossings, which tracked major shifts via the number of times an axis crossed from positive to negative. Every window of data was labeled with the majority movement state and kept if the movement state for 95% of the window's rows matched. This approach was taken to establish a threshold for noise tolerance to include more windows of transition states than if the threshold was set to 100% of the same class label, thus balancing this objective with ensuring minimal noise in the data.

Class imbalance across the five movement states was evident [Fig 1]. As mentioned previously, this was innate to the data given that these transitional states are brief and infrequent in the larger scheme of movement data collected in the WearGait PD study. Unfortunately, some data were lost due to dropping rows containing missing values, as there was not a consistent pattern of missingness, but rather missing data was spread across 752 of the 935 features. The split between PD patients and controls was still split about evenly, even after dropping rows with missing data [Fig 2]. Future renditions of this project may consider imputation as a means to retain more of the available data.

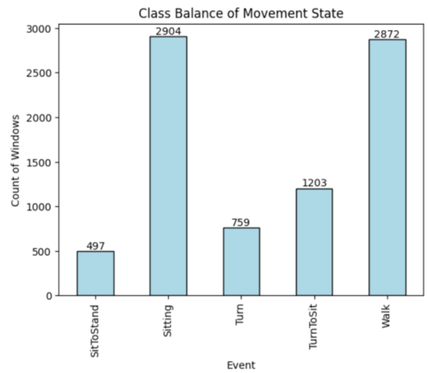


Fig 1. Class imbalance of movement state.

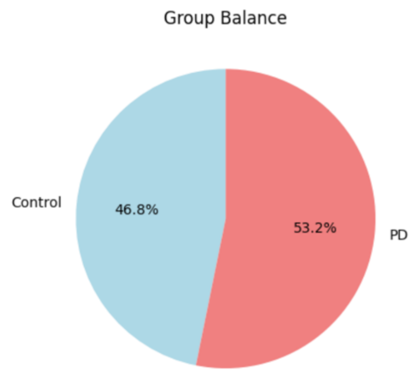


Fig 2. Groups were evenly represented in the data after pre-processing.

After feature engineering, the data was composed of 935 features and 8235 rows. A lasso logistic regression model was chosen to reduce dimensionality, while maintaining interpretability. The lasso regularization forces many weights to zero, thus eliminating redundant or uninformative features, and providing insight into which sensor locations are most useful in predicting movement states. The regularization strength was selected using Group-10-fold cross validation. The best performing model set the regularization parameter equal to 0.01, a relatively small value. Given the parameter's inverse relationship with regularization, a small parameter value leads to a high degree of regularization, with many coefficients pushed to zero, as is evident in the feature space being reduced from 935 to 164 features.

A second model was developed to address the first aim of predicting movement states. The data had high dimensionality in which multiple features were derived from the same sensors, thus resulting in multicollinearity. Principal components analysis (PCA) was chosen to address colinear features. PCA identifies the directions along which the variance in the data is maximized. The proportion of variance was analyzed [Fig 3.] and 278 components were selected, which accounted for 97% of the variance. This selection was done prior to running the model via cross-validation to improve computational efficiency.

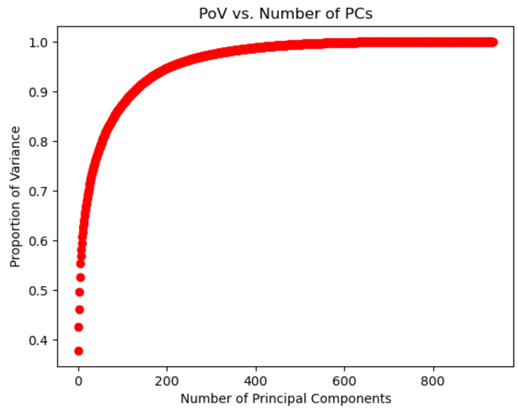


Fig 3. Proportion of variance captured by the number of principal components.

A support vector machine (SVM) model was implemented via sci-kit-learn's GridSearchCV with 5-folds and a gaussian/radial-basis function (RBF) kernel. SVM was selected for this task due to its ability to map data to a non-linear, higher-dimensional feature space. In essence, the pipeline selected 278 components to represent the original feature space. Then, the RBF kernel mapped these components into a higher-dimensional feature space to uncover non-linear relationships between samples. GridSearchCV was used to select the optimal parameter C and gamma that maximized model accuracy [5].

C. Aim 2 – Feature Engineering and Model Development

Data Preprocessing

The second aim of predicting Parkinson's Disease status again involved the derivation of windowed data. However, rather than compute summary statistics per window to represent meaningful features, a convolutional neural network was chosen for its ability to derive features via convolved kernels traversing the feature space. Windows of 2-seconds were first explored, and then increased to 5-second windows which achieved better results as it better captured long-duration symptoms such as gait impairment and bradykinesia (slowness of motion). Additional data was included for the second aim, because once the window size was expanded this naturally led to a fewer number of samples. As such, raw data from the self-paced and hurried-pace walking tasks and well as the self-paced turning task were included to augment the size of the dataset. The walking tasks were selected for data augmentation because of the frequent occurrence of these motions in every-day life, ensuring that the model would be well-equipped to detect Parkinsonian symptoms in the context of every-day movement patterns. The turning task was included to increase the number of transitional states in the dataset. Given the gait impairment characteristic of PD, complex transitional movements are likely a less fluid movement than when executed by unaffected control subjects, and thus could be an important feature used to identify PD. In sum, the final rendition of the preprocessed data included 5-second sliding windows with 20% overlap.

After extracting the windowed data, the model was trained using GroupKFold cross-validation with 5-folds on 80% of the data, with the remaining 20% set aside as a hold-out set, so that the model could be evaluated on unseen data. GroupKFold was chosen to ensure that a single participant's data, which accounted for multiple windows, was not separated between the training and testing sets. These measures were taken to prevent data leakage. Within each fold, the training and validation data were scaled using sci-kit-learn's StandardScaler. Prior to evaluation on the holdout set, the holdout set was scaled with the scaler first being fit on the training data.

A convolutional neural network was chosen for to address aim 2 due to its unique capability of extracting meaningful features via the convolutional layers. The convolutional layers were designed based on knowledge of Parkinsonian symptoms. The first layer employed a smaller kernel which traversed quarter-second intervals within each window in order to detect fine-grain and short-duration details such as the Parkinsonian tremor or a brief hesitation when initiating movement. A second convolutional layer employed a mid-sized kernel traversing 2-seconds of the 5-second window, detecting abnormalities during transitional movements. Finally, the third convolutional layer was intended to examine the entire 5-minute window with a large kernel, employing dilation rate to minimize computational complexity. This was with the intent of detecting symptomatology such as bradykinesia (slowness of motion) and gait impairment, which are only evident over a longer-duration window. Batch normalization and dropout layers were included after each convolutional layer to improve generalization and decrease overfitting. Early stopping with a patience of 5 epochs was another measure taken to prevent overfitting. Finally, global average pooling was used to flatten the output and passed it into the dense layers, which employed a softmax function for the binary class problem. The class distribution was approximately balanced (50.2% vs. 49.8%), so evaluation focused on accuracy and AUC rather than F1-measure.

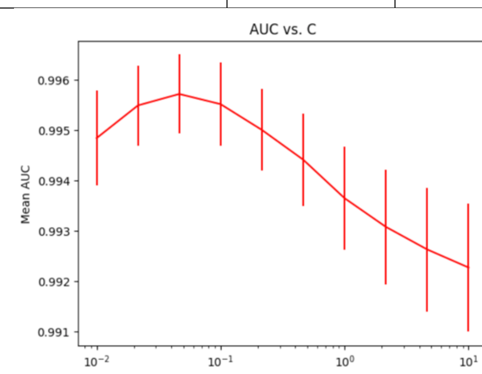
IV. Results and Discussion

A. Aim 1

The two models developed to address aim 1 performed comparably. Lasso logistic regression performed best with a small C parameter, translating to a large degree of regularization. [Fig 4.]

L1 Logistic Regression Results

	Metrics		
	Accuracy	***AUC***	***F1-measure***
L1 Logistic Regression	0.9475	0.9955	0.9818



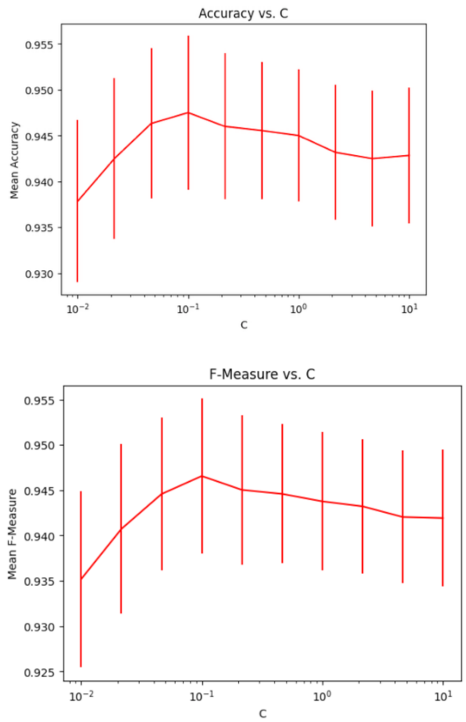


Fig. 4. Lasso Logistic Regression results

As shown in the charts in Figure 4., the optimal regularization parameter maximized all metrics: accuracy, AUC, and F1-measure. AUC was included to weigh false-positive and true-positive rates. F1-measure was included to assess how the model handled minority class prediction. As seen in the figures, the model performed well across all metrics.

Previous research by Anderson et al. investigated the importance of sensor location in detecting stride. The research team reported that the dorsal foot sensor achieved the best performance, followed closely by other lower limb sensors. This contrasted with the wrist sensor which performed significantly worse than all lower body sensors [7]. Their findings hold implications for the consumer wearables market as well as for PD detection, suggesting the greatest reliability of foot or leg-mounted sensors, in contrast to a consumer wearables market that is saturated with wrist sensors. Interestingly, the lasso logistic regression model selected features from sensors across the entire body, with the wrist sensor being selected the most [Fig 5]. It is important to note that Anderson et al. isolated each sensor in their analysis, providing a more robust answer to this particular question, whereas in the current study the model was optimized by using some but not all features from a sampling of locations across the entire body.

Moreover, Anderson et al. reported that complex or transitional actions such as moving from a sitting to a standing position posed a more difficult task for all sensors, with sensor-specific F1 scores as low as 50% and none above 90% [7]. The current study results validate these findings. It is clear that the brief and distinct transitional movement states captured in the TUG accounted for the majority of the error in the model [Fig. 6].

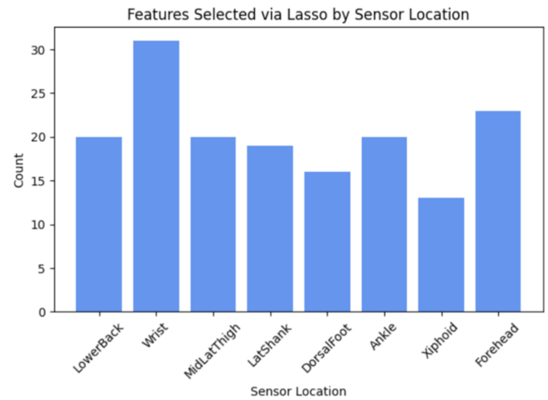


Fig 5. Sensor locations seleceted after lasso

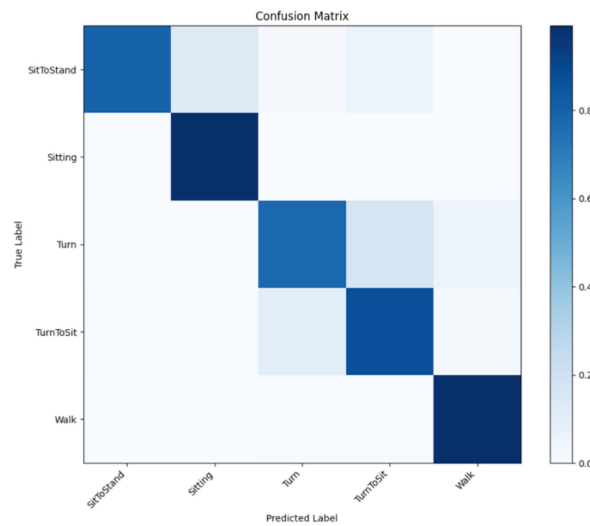


Fig. 6. Confusion matrix showing that transitional states accounted for the error.

The SVM model with a PCA pipeline performed as well as the lasso logistic regression model. However, due to the transformation of PCA it is less explainable, and thus less preferred. The best performing SVM model selected a high value of C for regularization strength, thus minimizing loss and misclassifications, and a low value of gamma, resulting in a wider kernel that is less concentrated in its emphasis of patterns among the samples [Fig 7]. The error in the SVM model is also largely accounted for by predictions on the transitional states.

SVM WITH RBF KERNEL RESULTS

	Metrics	
	Accuracy	***F1-measure***
SVM	0.9511	0.9505

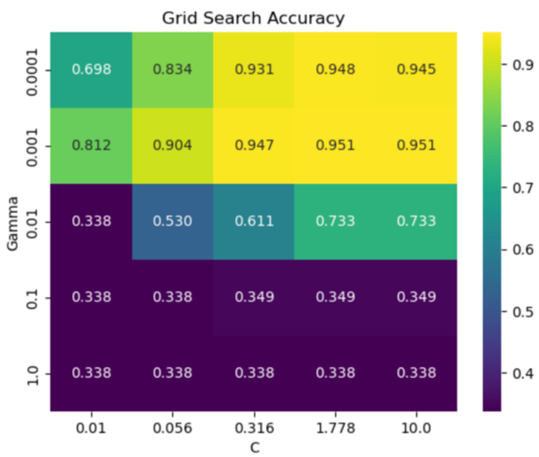


Fig. 7. Results of cross-validation in tuning SVM parameters.

B. *Aim 2*

A convolutional neural network (CNN) was developed to address the second aim of predicting PD status. The best performing model reported validation accuracy of 97% on the split on which it was trained. However, to fairly evaluate the model, the best performing model was applied to the holdout set, which accounted for 20% of all available preprocessed data. When run on the holdout set, the model reported an accuracy of 0.90 and an AUC of 0.954 [Fig 8].

CNN RESULTS

	Metrics	
	Accuracy	***AUC***
CNN	0.90	0.954

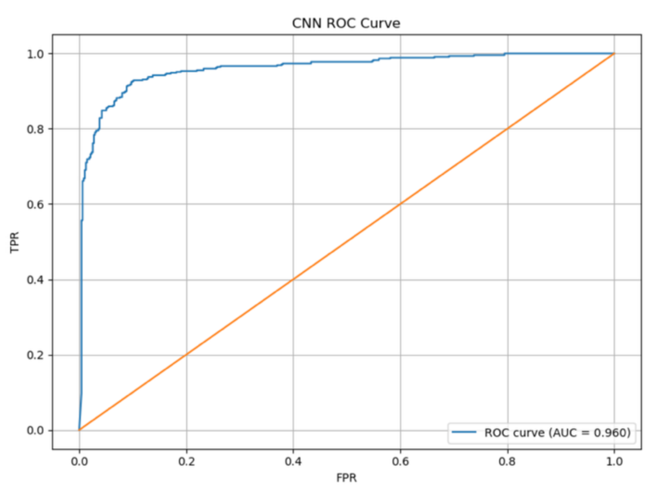


Fig. 8. Results of CNN showing AUC of receiver operating characteristic.

These results improve upon the existing literature. Schalkamp et al. explored a similar question pertaining to PD diagnostic capability using a large set of accelerometer data gathered by the UK Biobank. The team reported an AUC of 0.78 in distinguishing diagnosed PD with unaffected age-matched controls. Moreover, the lasso logistic regression model employed by Schalkamp et al. included age and sex as features in the model, whereas the current project's approach reports an even better AUC without the use of age and sex as covariates. The accelerometer data supplied by the UK Biobank was collected over the course of a day, and the researchers found that mean acceleration per hour was the most important factor in differentiating PD patients from controls [3]. Given that data used in the current project represents data collected over the course of minutes rather than hours, the results of the current project can be viewed within the context of solving a new problem, in addition to validating and furthering the current literature.

Similarly, the Parkinson's Disease Digital Biomarker DREAM Challenge crowdsourced approaches to predict PD status and symptom severity from a smartphone-based study and wrist accelerometer study, reporting an Area-under-the-Curve (AUC) of 0.87 for the model best at predicting PD status. As reported by the DREAM challenge, the top performing submissions included lasso logistic regression models using summary statistics of windowed data, signal processing techniques namely Fourier Transform, the use of the python package tsfresh to extract features from raw IMU data, and the development of deep-learning models, namely CNNs [2].

V. CONCLUSION AND FUTURE DIRECTIONS

The current study accomplished two aims.

1) Predict the five movement states present in the TUG task, including complex transitional states, to assess the efficacy of IMU wearables in classifying distinctive motions. The lasso logistic regression model and gaussian kernel SVM models both performed well, handling class imbalance gracefully with F1 measures just as elevated as accuracy and AUC scores. The error on these models can be largely explained by the minority class: the disctinctive transitional movements. Future directions could employ data augmentaiton to replicate the protocol and increase the number of samples representing specific transitional states like turning to sit in a chair to achieve better results. Overall, the current project establishes the high reliability and accuracy of IMU sensor data in interpreting specific motion patterns. This provides support for clinical applications of IMU sensors to monitor health, as the renderings of such models must be hyper-precise in order to be relied on for health tracking.

2) Discern PD status from a balanced group of PD patients and age-matched controls, to establish the feasibility of IMU sensors for disease monitoring and detection. A CNN was thoughtfully designed to address this task, with convolutional layers selected to detect specific Parkinsonian symptoms. The results of the CNN model are promising, validating the use of deep-learning techniques for clinical applications. The model's robust results would be improved with more data and perhaps additional convolutional layers. Overall, IMU wearable devices present an accessible and cost-effective method of detecting PD via a data source that is becoming increasingly ubiquitous with the advent and growth of the consumer wearables market.

References

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *J. Neurol. Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] S. K. Sieberts, J. Schaff, M. Duda, et al., "Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge," *npj Digital Medicine*, vol. 4, no. 53, 2021. [Online]. Available: https://doi.org/10.1038/s41746-021-00414-7

[3] A. K. Schalkamp, K. J. Peall, N. A. Harrison, *et al*., "Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis," *Nature Medicine*, vol. 29, pp. 2048–2056, 2023. [Online]. Available: https://doi.org/10.1038/s41591-023-02440-2

[4] R. Piwek, D. A. Ellis, S. Andrews, and A. Joinson, "The rise of consumer health wearables: Promises and barriers," *PLOS Med.*, vol. 13, no. 2, p. e1001953, 2016. [Online]. Available: https://doi.org/10.1371/journal.pmed.1001953

[5] A. J. Anderson, D. Eguren, M. A. Gonzalez, et al., "WearGait-PD: An open-access wearables dataset for gait in Parkinson's disease and age-matched controls," *medRxiv*, preprint, 2024. [Online]. Available: https://doi.org/10.1101/2024.09.11.24313476.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] A. J. Anderson, M. Gonzalez, D. Eguren, et al., "Deep learning-based **stride segmentation** with wearable sensors: effects of data quantity, sensor location, and task," *medRxiv*, preprint, 2024. [Online]. Available: https://doi.org/10.1101/2024.12.17.24319160.