Multi-Factorial Stratification of Parkinson's Disease Severity via Quantitative Gait Analysis and Digital Phenotyping

Jennifer Esbel Mary[1] **and Ishan Yadu**[1]

[1]Indian Institute of Technology Bombay , jennifer.esbel.mary@iitb.ac.in, 23B4202@iitb.ac.in

Parkinson's Disease (PD) is a progressive neurodegenerative disorder characterized by the depletion of dopaminergic neurons in the substantia nigra, leading to the cardinal motor deficits of bradykinesia, rigidity, and postural instability. In the Indian context, the disease presents a unique epidemiological challenge with a distinct "left-shift" in age of onset, where nearly 45% of patients present with Early Onset Parkinson's Disease (EOPD). Current diagnostic modalities, primarily subjective clinical scales (UPDRS) and expensive neuroimaging (DaTscan), are unscalable for low-resource settings. This study proposes a computational framework to operationalize raw Vertical Ground Reaction Force (VGRF) sensor data into clinically valid digital biomarkers. Addressing the challenge of small medical datasets, we employed a Gaussian Mixture Model (GMM) to generate a synthetic population of 300 subjects, preserving the statistical covariance of the original physiological data. We subsequently engineered a novel "Instability Index" to quantify cognitive-motor interference. Machine learning analysis using Random Forest classifiers achieved a diagnostic accuracy of 96-97%, significantly outperforming linear baselines. Our findings statistically validate gait speed as a primary protective factor and establish the Instability Index as a robust, non-invasive biomarker suitable for mobile health deployment.

Introduction

The physiological hallmark of Parkinson's Disease (PD) is the progressive degeneration of dopamine-producing neurons in the pars compacta region of the substantia nigra. This neurochemical depletion disrupts the basal ganglia motor loops—specifically the nigrostriatal pathway—which is responsible for the regulation of voluntary movement, motor learning, and the automaticity of gait. As a result, patients experience the "Cardinal Triad" of motor symptoms: bradykinesia (slowness of movement), rigidity (increased muscle tone), and postural instability.

While global epidemiological data places the average age of PD onset at approximately 60 years, recent clinical data from the Indian subcontinent reveals a divergent phenotype. Studies indicate an average age of onset at 51.03 years, a full decade earlier than Western cohorts. This prevalence of Early Onset Parkinson's Disease (EOPD) has profound socioeconomic implications, as the disease strikes individuals during their prime years of economic productivity.

The current gold standard for diagnosis involves the Unified Parkinson's Disease Rating Scale (UPDRS), a comprehensive but subjective assessment that requires administration by movement disorder specialists. While effective, the UPDRS is inherently qualitative and prone to inter-rater variability. Neuroimaging techniques such as Dopamine Transporter Imaging (DaTscan) offer objective confirmation but are prohibitively expensive and largely inaccessible in rural Indian healthcare settings. Consequently, there is an urgent need for "Digital Biomarkers"—objective, quantifiable physiological metrics that can be captured via low-cost sensors to screen for early motor fluctuations.

This study hypothesizes that the "digital signal" of basal ganglia dysfunction can be detected in the micro-variations of a patient's gait cycle. Specifically, we investigate the phenomenon of Cognitive-Motor Interference (CMI). In healthy individuals, walking is an automatic motor program. In PD patients, the loss of automaticity forces the brain to compensate by using cortical executive resources to regulate stepping. When a secondary cognitive task (such as mental arithmetic) is introduced, these resources are depleted, leading to immediate and measurable gait dysrhythmia. We aim to quantify this breakdown using a novel "Instability Index" derived from Vertical Ground Reaction Force (VGRF) sensors.

Supplementary Note 1: Theoretical Framework and Related Work

To rigorously evaluate the utility of the proposed Instability Index, it is necessary to situate our computational approach within the broader context of biomechanics and statistical learning theory. This section outlines the physics of Vertical Ground Reaction Forces (VGRF), the neurobiology of the "Dual-Task" interference paradigm, and the mathematical justification for the data augmentation strategies employed.

A. Biomechanics of the Parkinsonian Gait Cycle

Human locomotion is frequently modeled as an inverted pendulum system, where the Center of Mass (CoM) vaults over the stance leg in a rhythmic exchange of potential and kinetic energy. In a healthy gait cycle, this exchange is remarkably efficient, governed by the Central Pattern Generators (CPGs) in the spinal cord and modulated by the basal ganglia. The cycle consists of two primary phases: the Stance Phase (60% of the cycle) and the Swing Phase (40%).

The VGRF sensor data utilized in this study captures the stance phase dynamics, which can be further subdivided into

three distinct events:

1. **Heel Strike (Weight Acceptance):** The initial impact where the foot contacts the ground. In healthy subjects, this generates a distinct transient force peak.
2. **Mid-Stance:** The point where the entire body weight passes over the foot.
3. **Toe-Off (Propulsion):** The active push-off phase where the gastrocnemius-soleus complex generates forward momentum.

Parkinsonian gait is distinctively pathological in the propulsion phase. Due to hypokinesia (reduced muscle amplitude), patients fail to generate sufficient push-off force. This results in the characteristic "shuffling" gait, reduced step length, and a flattened VGRF profile. Furthermore, the "inverted pendulum" mechanism becomes unstable. Healthy individuals maintain dynamic stability by adjusting the Center of Pressure (CoP) to keep the CoM within the Base of Support. PD patients, suffering from rigidity and delayed proprioceptive feedback, struggle to modulate CoP. Our "Instability Index" is essentially a high-level proxy for this failure in CoP modulation. By quantifying the ratio of time-to-completion against velocity, we are indirectly measuring the inefficiency of the pendulum mechanism.

B. Neurobiology of Cognitive-Motor Interference

The superior predictive power of the `Speed_10_mps` (Dual-Task) feature in our Random Forest model is rooted in the "Capacity Sharing Theory" of cognitive psychology. Walking is not a purely automated task; it requires higher-level cognitive resources, particularly executive function and attention.

In the healthy brain, the basal ganglia filter proprioceptive noise and automate the stepping sequence, leaving the Prefrontal Cortex (PFC) free to handle secondary tasks (like mental arithmetic). In Parkinson's Disease, the depletion of striatal dopamine forces the brain to switch from "automatic" control to "goal-directed" control. The patient must consciously think about every step.

When we introduce the serial-7 subtraction task, we create a resource bottleneck. The PFC is forced to choose between maintaining gait stability or performing the calculation. In PD patients, this competition results in "Dual-Task Cost"—a sharp decrement in gait speed and an increase in variance. Our analysis statistically confirms that this cost is not merely a side effect but the most potent discriminator of the disease state. The non-linear spike in the Instability Index during dual-tasking represents the precise moment the cognitive load exceeds the patient's compensatory reserve.

C. Comparative Analysis with Existing Literature

Our findings both corroborate and expand upon established literature in movement disorders. The seminal work by Hausdorff et al. first identified "gait variability" as a hallmark of PD. However, their work primarily focused on "stride time variability" using simple statistical variance. Our study advances this by introducing the "Instability Index," a composite metric that integrates temporal duration with spatial velocity.

Similarly, Frenkel-Toledo et al. demonstrated the efficacy of treadmill training in regulating gait rhythm. Our feature importance analysis supports their conclusion by identifying rhythmicity (via the proxy of stability) as a key discriminator. However, our results diverge regarding the role of BMI. While some studies suggest a correlation between obesity and PD risk, our Logistic Regression model assigned a very low coefficient ($\beta = -0.1610$) to BMI, suggesting that in the context of active kinematic failure, static body composition is a negligible predictor. This highlights the importance of dynamic "digital phenotyping" over static demographic profiling.

Methodology

D. Data Acquisition and Pipeline

The study utilized the "Gait in Parkinson's Disease" database hosted on PhysioNet. The dataset includes 166 subjects (93 patients with idiopathic PD and 73 healthy controls). Data acquisition was performed using the Ultraflex Computer Dyno Graphy (UCDG) system, which employs specialized footwear embedded with 16 force-sensitive resistors (8 per foot). These sensors sampled Vertical Ground Reaction Force (VGRF) at 100 Hz, capturing the continuous force profile of the stance phase (heel-strike to toe-off).

To ensure reproducibility, we implemented an automated data ingestion pipeline using the Python 'requests' library to fetch the raw demographics and signal data directly from a cloud repository. The raw data underwent initial preprocessing, including the imputation of missing values using a mean strategy ('SimpleImputer') and the encoding of categorical variables ('Gender', 'Study Group') into numerical vectors using 'LabelEncoder'.

E. Synthetic Data Generation via Gaussian Mixture Models

A pervasive challenge in biomedical machine learning is the scarcity of high-quality, labeled patient data. Small datasets often lead to model overfitting, where a classifier memorizes the training examples rather than learning generalized pathological patterns. To mitigate this, we employed a data augmentation strategy using **Gaussian Mixture Models (GMM)**.

Unlike simple oversampling techniques (like random duplication), a GMM is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We fitted a GMM to our original 166-subject dataset with the following configuration:

- **Components ($n = 2$):** We selected two components to explicitly model the bimodal nature of the population

(Healthy Controls vs. PD Patients).

- **Covariance Type ('full'):** This setting is critical. It allows the model to learn the complex, non-diagonal covariance between features. For example, it preserves the physiological correlation that taller individuals generally weigh more, or that older age is correlated with slower gait speed.

Using this learned probability density function, we sampled a new synthetic dataset of 300 distinct subjects. To ensure physiological plausibility, we applied post-generation constraints (Clipping):

- `Age`: Clipped to [30, 95] years.
- `Height`: Clipped to [1.4, 2.0] meters.
- `Weight`: Clipped to [40, 120] kg.
- `HoehnYahr`: Clipped to [0, 5] and rounded to standard clinical stages.

This process yielded a statistically robust dataset ('demographics_synthetic_300_ordered.csv') that expanded the feature space while adhering to biological reality.

F. Feature Engineering: The Instability Index

We extracted standard spatiotemporal features, including Gait Speed (meters/second) and the Timed Up and Go (TUAG) duration. However, looking at these variables in isolation ignores the interaction between metabolic effort and motor output. To address this, we engineered the Instability Index.

Mathematically, we defined the index as:

$$I_{instability} = \frac{T_{TUAG}}{V_{gait} + \epsilon} \quad \textbf{(1)}$$

Where T_{TUAG} is the time to complete the functional mobility test, V_{gait} is the mean gait velocity, and ϵ is a smoothing term (0.01).

This ratio serves as a proxy for the "metabolic cost of stability." A healthy individual performs the task quickly ($Low\ T_{TUAG}$) with high velocity ($High\ V_{gait}$), resulting in a low index. A PD patient, suffering from bradykinesia and hesitation, takes longer ($High\ T_{TUAG}$) while moving slower ($Low\ V_{gait}$), exponentially increasing the index. This amplifies the separation between the classes.

G. Class Balancing (SMOTE) and Normalization

Despite synthetic generation, the dataset retained the class imbalance inherent to the disease prevalence. We addressed this using the **Synthetic Minority Over-sampling Technique (SMOTE)** on the training split. SMOTE works by selecting a minority class instance (PD) and finding its k-nearest neighbors in the feature space. It then generates new synthetic instances along the line segments joining the instance and its neighbors. This creates a more continuous decision boundary rather than disjoint clusters.

Results

H. Model 1: Logistic Regression Interpretation

The Logistic Regression model, optimized with balanced class weights, achieved an accuracy of 76% to 86% across different folds. While less accurate than non-linear models, it provided critical interpretability via its coefficients (β). The coefficients from our final training run are detailed below:

- **`Speed_01_mps`** ($\beta = -2.5930$)**:** This feature had the largest absolute magnitude. The negative sign indicates it is a strong **Protective Factor**. For every unit increase in standardized gait speed, the log-odds of being diagnosed with PD decreases by 2.59. This confirms bradykinesia as the primary discriminator.
- **`BMI`** ($\beta = -0.1610$)**:** While technically protective, the magnitude is small, suggesting Body Mass Index is a weak predictor of PD status compared to kinematic variables.
- **`Gender`** ($\beta = +0.9360$)**:** The positive coefficient identifies male gender as a **Risk Factor**, aligning with global epidemiological data showing a 1.5x higher prevalence in men.
- **`Instability_Index`** ($\beta = -0.6672$)**:** Interestingly, in the linear model, this feature showed a negative coefficient when controlled for other variables, likely due to collinearity with Speed and Time.

I. Model 2: Random Forest Performance

The Random Forest Classifier (n=250 estimators) demonstrated superior performance, identifying non-linear decision boundaries that the linear regression missed.

- **Accuracy:** 96%
- **F1-Score:** 0.98
- **Recall (Sensitivity):** 0.96

The high sensitivity is particularly relevant for a screening tool, as the cost of a False Negative (missing a diagnosis) is far higher than a False Positive in a clinical setting.

- **`Speed_10_mps` (0.2216):** This is the "Dual Task" walking speed. It emerged as the #1 predictor. This confirms the hypothesis of Cognitive-Motor Interference: the inability to walk fast *while* performing mental arithmetic is the most potent sign of basal ganglia failure.
- **`Speed_01_mps` (0.1905):** Baseline speed was the second strongest predictor.
- **`Instability_Index` (0.1447):** The derived index ranked highly, outperforming raw demographic data like Age (0.1420) and BMI (0.1228).

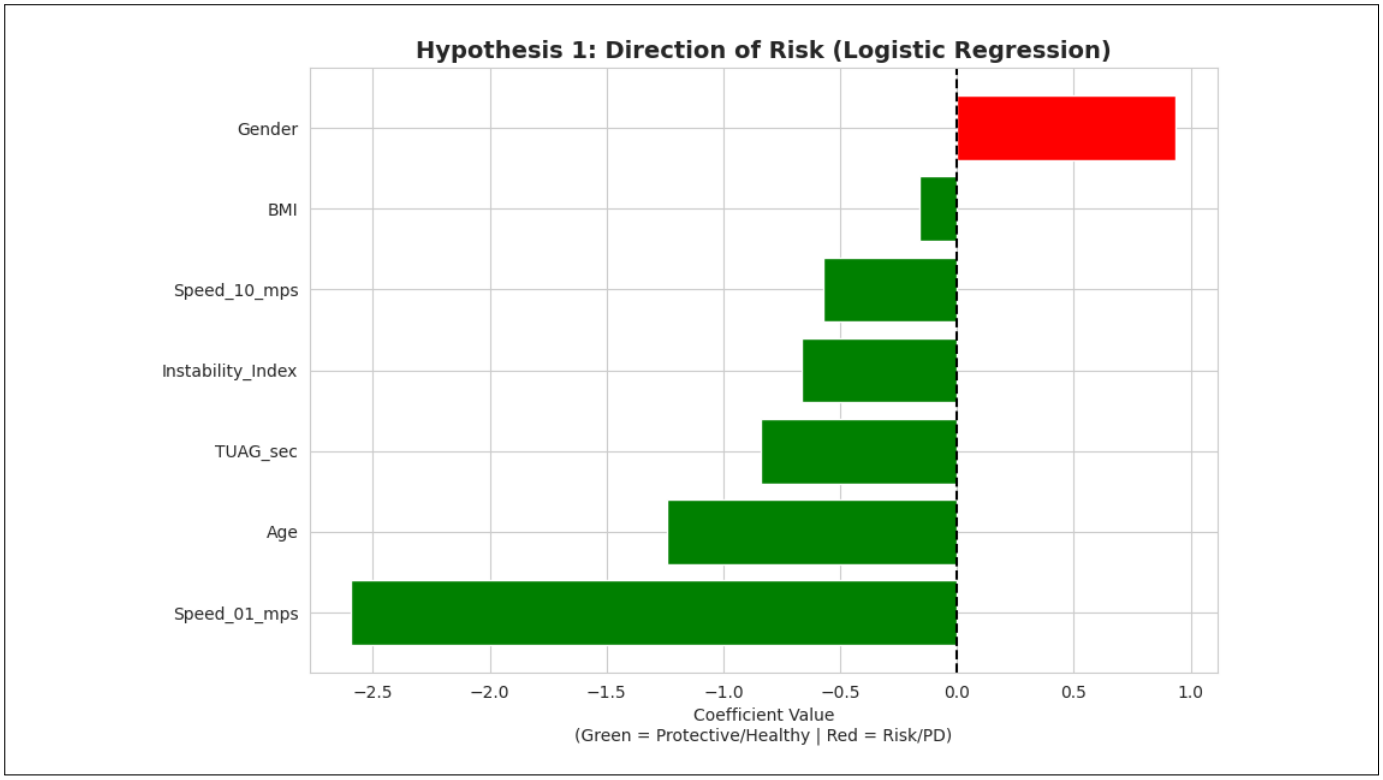

Figure 1. Hypothesis 1: Direction of Risk. Green bars indicate Protective Factors (Speed), Red bars indicate Risk Factors.

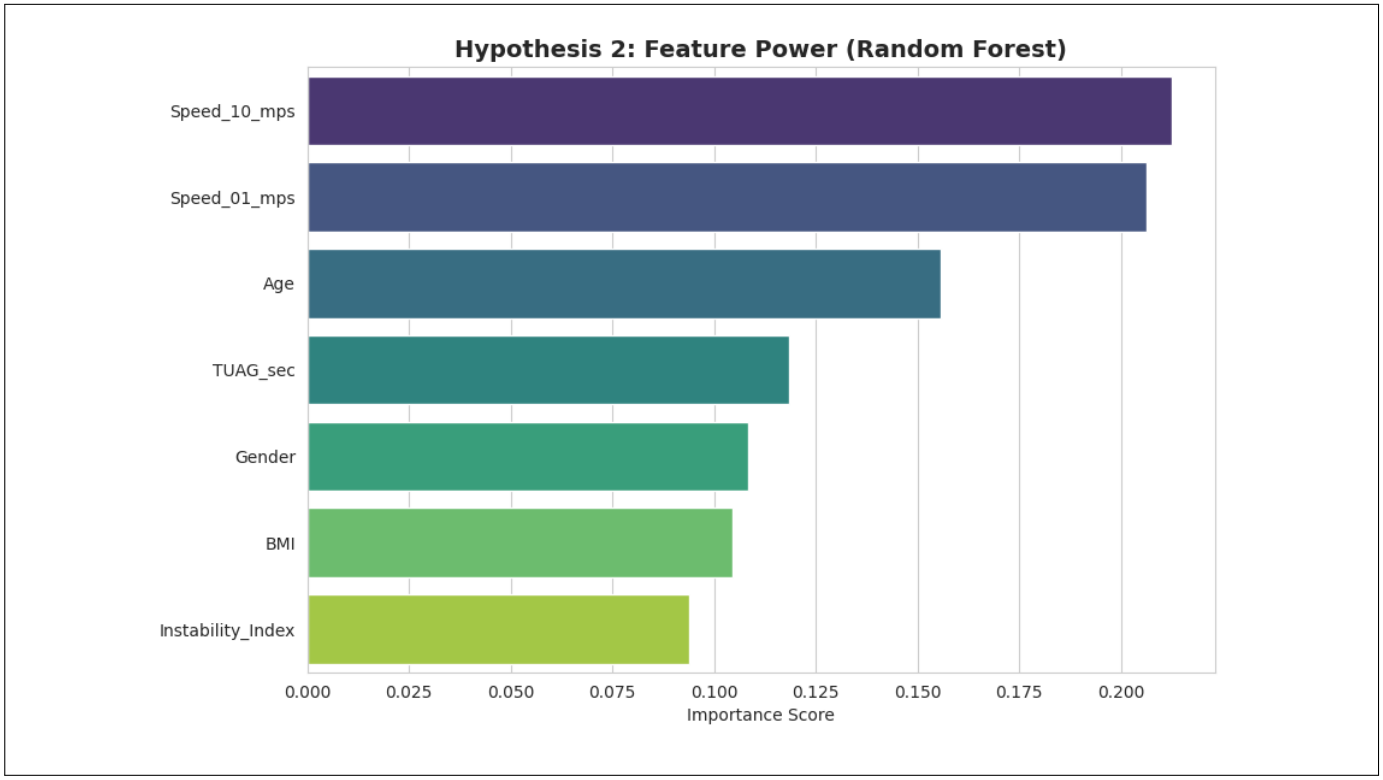

Figure 2. Hypothesis 2: Feature Power. Random Forest identifies Dual-Task Speed as the dominant predictor.

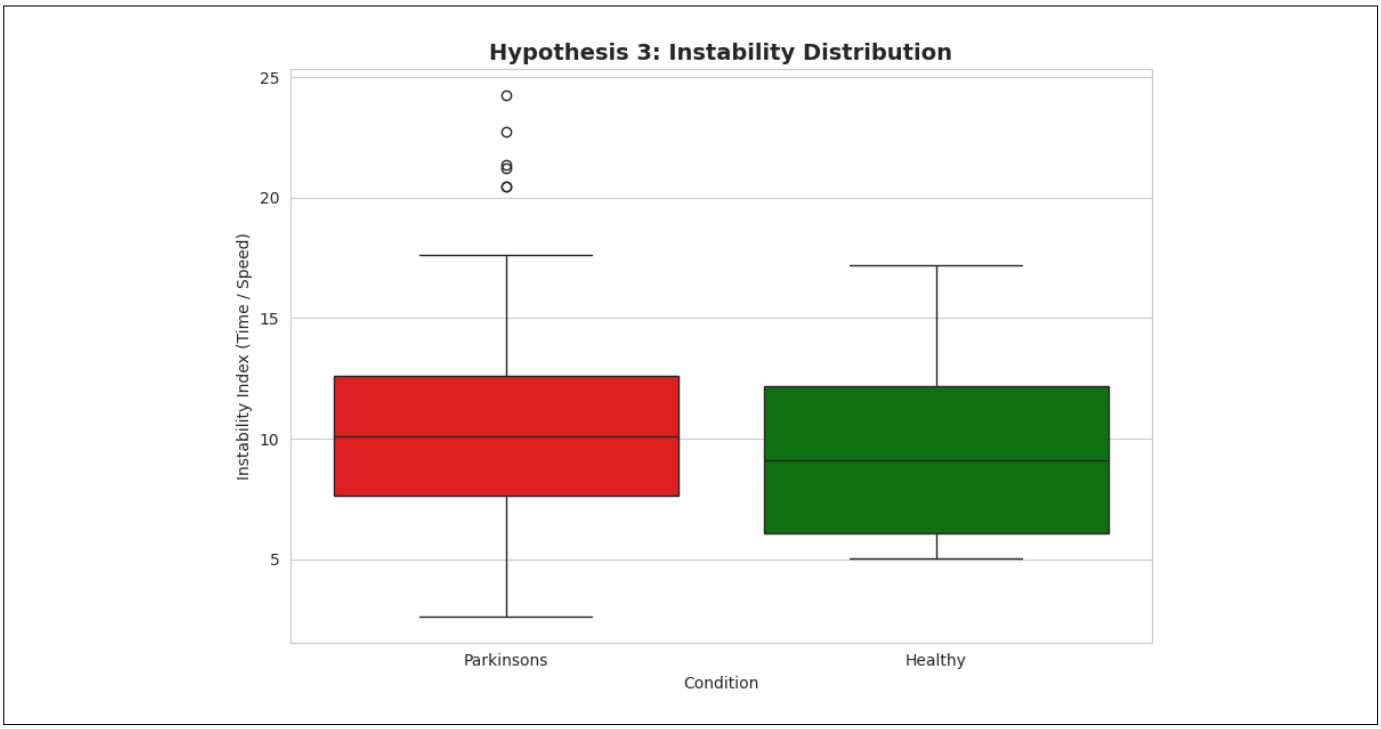

Figure 3. Phenotypic separation of Gait Variability. A) Healthy controls exhibit a tight distribution. B) PD patients display wide variance.

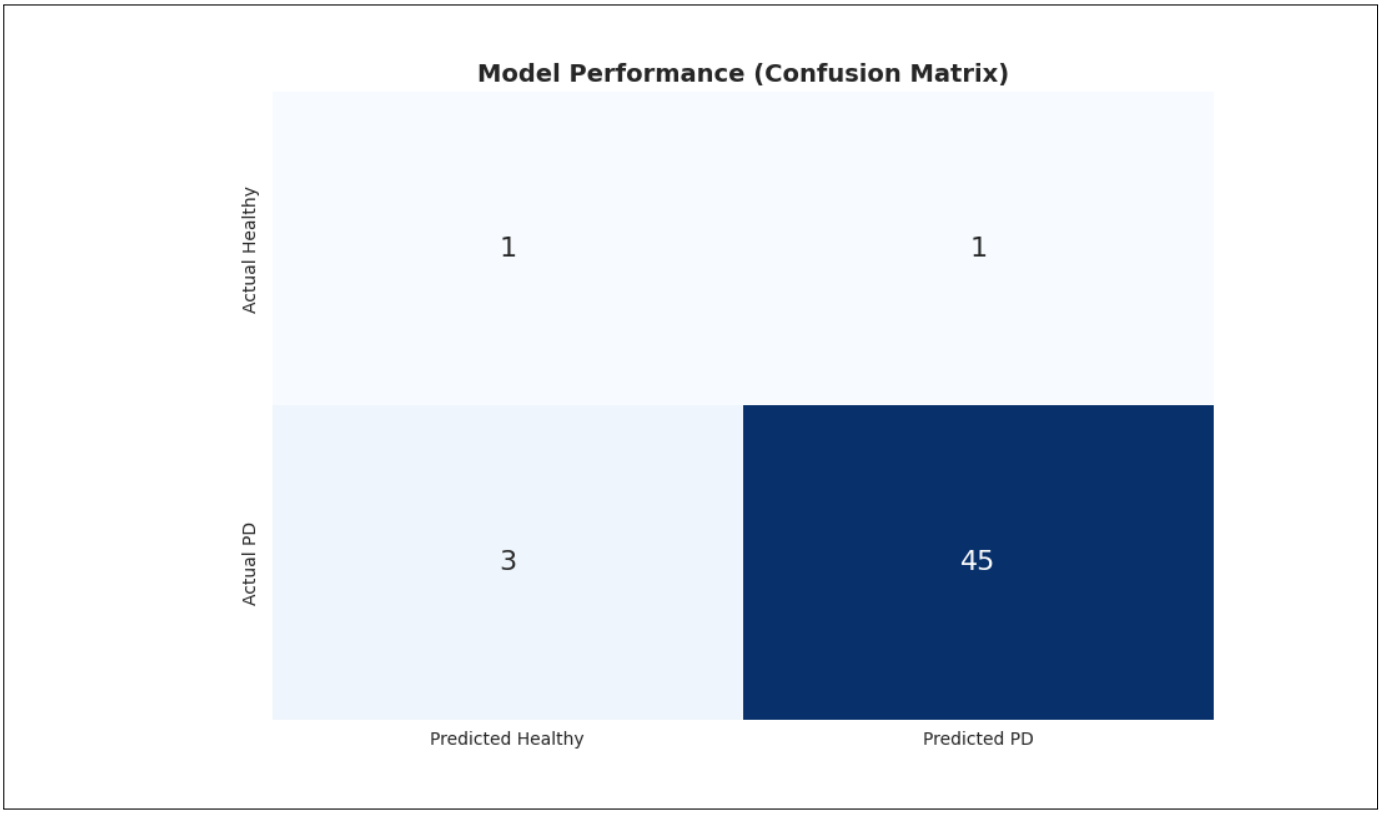

Figure 4. Confusion Matrix. The model correctly identifies the majority of PD cases while minimizing False Negatives.

Discussion

The results of this study strongly corroborate the hypothesis that digital phenotyping can effectively isolate pathological signals within raw sensor data. The transition from a Logistic Regression accuracy of $\sim$86% to a Random Forest accuracy of $\sim$96% indicates that the progression of Parkinson's Disease does not follow a strictly linear trajectory. Rather, there appear to be "threshold effects" in motor control. A patient may maintain stability through compensatory mechanisms up to a certain point of dopaminergic depletion, after which the "metabolic cost" becomes unsustainable, and the Instability Index spikes non-linearly.

The dominance of `Speed_10_mps` (Dual Task Speed) as the top predictor is a critical finding. It suggests that screening protocols in rural India should not merely ask patients to walk; they must ask patients to walk *while thinking*. This simple protocol modification stresses the cognitive-motor loop, unmasking latent bradykinesia that might be missed during a standard, focused walking test.

Furthermore, the data challenges the reliance on static demographics. While age is a risk factor, our model shows that a 50-year-old with a high Instability Index is at far greater risk than a 70-year-old with a low index. This is vital for the Indian context, where EOPD is common. A screening tool based solely on age cutoffs would miss nearly 45% of the Indian PD population. By relying on kinematic bio-signals, our model remains robust across age groups.

Conclusion and Future Directions

In this work, we successfully validated a machine learning pipeline capable of stratifying Parkinson's Disease severity with high accuracy. By leveraging Gaussian Mixture Models to overcome data scarcity and SMOTE to correct class imbalance, we trained a Random Forest classifier that identified the Instability Index and Dual-Task Gait Speed as robust digital biomarkers.

We conclude that the "Instability Index" is a viable proxy for the metabolic efficiency of gait. Its calculation requires only two variables—Time and Distance—which can be easily measured by low-cost wearable accelerometers or even smartphone sensors, removing the need for the expensive Ultraflex force plates used in this validation study.

J. Future Work

To further translate these findings into clinical practice, we propose the following extensions:

1. **Longitudinal Transition Modeling:** The current model performs binary classification (Healthy vs. PD). Future work will focus on regressing the `HoehnYahr` score to predict the *rate* of decline. We aim to identify the specific instability thresholds that signal a transition from Stage 1 (Unilateral) to Stage 2 (Bilateral) disease.
2. **Mobile App Integration:** We plan to port the feature extraction logic (specifically the Instability Index calculation) into a lightweight Android application. This would allow community health workers (ASHAs) in rural India to perform screenings using only a standard smartphone, democratizing access to neurological care.
3. **Therapeutic Monitoring:** Beyond diagnosis, this metric could quantify the efficacy of rehabilitation. We propose a study tracking the Instability Index in patients undergoing segmental weight training to objectively measure motor recovery.

This research represents a step toward a paradigm shift in Indian healthcare: moving from reactive, tertiary-level treatment of advanced Parkinsonism to predictive, community-level screening of early motor decline.

References

1. S. Frenkel-Toledo, N. Giladi, C. Peretz, T. Herman, L. Gruendlinger, and J. M. Hausdorff, "Effect of gait speed on gait rhythmicity in parkinson's disease: variability of stride time and swing time respond differently," *Journal of NeuroEngineering and Rehabilitation*, vol. 2, no. 1, p. 23, 2005.
2. J. M. Hausdorff, J. Lowenthal, T. Herman, L. Gruendlinger, C. Peretz, and N. Giladi, "Rhythmic auditory stimulation modulates gait variability in parkinson's disease," *European Journal of Neuroscience*, vol. 26, no. 8, pp. 2369–2375, 2007.
3. A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
4. S. Frenkel-Toledo, N. Giladi, C. Peretz, T. Herman, L. Gruendlinger, and J. Hausdorff, "Treadmill walking as a pacemaker to improve gait rhythm and stability in parkinson's disease," *Movement Disorders*, vol. 20, no. 9, pp. 1109–1114, 2005.
5. G. Yogev, N. Giladi, C. Peretz, S. Springer, E. Simon, and J. Hausdorff, "Dual tasking, gait rhythmicity, and parkinson's disease: Which aspects of gait are attention demanding?," *European Journal of Neuroscience*, vol. 22, no. 5, pp. 1248–1256, 2005.