ParkinAid: A Multimodal System for Parkinson's Diagnosis through Motor and Speech Disorder Analysis

Liyang Han, 16900 West Gebhardt Road, Brookfield, WI 53005

Abstract

Early diagnosis of Parkinson's Disease (PD) is critical for effective intervention. However, traditional diagnostic methods rely on in-person consultations with specialists and expensive clinical evaluations, making them both costly and time-consuming. Existing software-based diagnostic tools often focus on a limited subset of physiological symptoms and are not widely accessible to the public.

ParkinAid is a web and app-based software system that leverages AI-driven computer vision and speech analysis models to predict Parkinson's-related index scores and perform binary classification based on motor and speech disorder assessments using UPDRS-III criteria. The system analyzes 18 features using 25 AI models, trained on over 133 video samples and 30 voice recordings. In body motion analysis, an RNN with GRU outperformed other models, achieving 90% accuracy and an MSE of 0.018. For hand tremor analysis, the NN-LSTM (Bi-Directional) model achieved the highest accuracy of 91.2%, precision of 90.5%, and recall of 92.0%. In speech analysis, the best-performing models for phonation and articulation were CRNNs (Stacked GRUs) and BERT (Large Fine-Tuned on Stress), achieving 89.5% and 93.1% accuracy, respectively.

ParkinAid has demonstrated high accuracy, robustness, and reliability in motor and speech disorder analysis. As the first system to standardize PD diagnosis under a unified scale, it provides an extensible platform for future developers to integrate and refine additional AI-driven diagnostic methods.

Key Words: Parkinson's, Diagnosis, Index Predict, Classification, Multi-modal, Computer vision, Web-based, Mobile App

Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the successful completion of this research and paper. First and foremost, I am deeply indebted to my computer science teacher Ryan Osterberg, whose guidance, support, and expertise were invaluable throughout the entire research process. Additionally, I extend my thanks to lots of previous researchers who have contributed to the Parkinson's field providing high-quality research, logs, papers related to the technology I referenced in my study. Finally, I would like to acknowledge the emotional and financial support received from my parents Lisa Wang and Jonathan Han that facilitated this study.

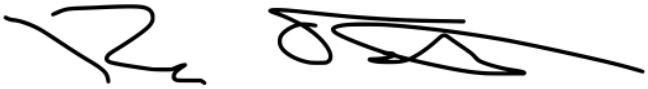
Teacher Sign: 

Table of Contents

1. Introduction

1.1 Background

Parkinson's disease (PD) is one of the most common neurodegenerative diseases. It leads to significant limitations in the daily life of patients over time. Early diagnosis of Parkinson's is crucial for effective management and improving quality of life. However, traditional diagnostic methods often require in-person consultations with specialists and expensive equipment, creating barriers for timely intervention. We developed a device in 2023 that provides the capability of diagnosis and rehabilitation aid to users at home [1], but it's a dedicated device and not convenient enough or widely available. Software diagnosis tool can provide a more cost-effective way and increase the access significantly.

1.2 Literature Research

The software diagnosis is based on physiological symptoms. The dominant modern medical standard for the diagnosis and assessment of the severity of Parkinson's disease symptoms is the MDS-UPDRS (Unified Parkinson's Disease Rating Scale) introduced in the 1980s [2], which includes motor symptoms and non-motor symptoms for PD diagnosis and rating of severity, including but not limited to bradykinesia, rigidity, resting tremor, and gait disturbances, olfactory impairment, orthostatic hypotension, constipation, sleep disturbances, and speech impairment[3][4][5].

Recent software diagnosis tool analyzes audio and videos to diagnose PD from multiple aspects[6][7][8]. Some studies attempted automated speech analysis [9]. Some developed machine learning models to assess PD by looking at motor symptoms and analyzing facial expressions and

movement gait disorders such as rigidity or tremor from recorded videos [10][11]. AI-driven analysis of speech and motion presents a promising avenue for the early detection of Parkinson's disease. But these works are mostly limited to a few aspects, but Parkinson is very complicated and show many symptoms. Meanwhile, although there are few online software tool [12] available, they most are not accessible to the public.

1.3 Solution

This work integrates motor symptoms and speech features to provide a more comprehensive assessment, potentially enhancing diagnostic accuracy and accessibility. It makes improvements to and introduces novel algorithms for movement and voice analysis models, designed to improve diagnostic precision. Additionally, the system increases accessibility for Parkinson's disease screening by offering a web-based and mobile application. This tool is particularly beneficial for individuals in rural areas with limited access to neurologists, providing a preliminary screening assessment for Parkinsonism.

ParkinAid is the name of the system created and will be referred to throughout this entire paper. ParkinAid's methods of diagnosing include:

A. Motor analysis, including body motion and hand tremor analysis, utilizes video input to predict various Parkinson's-related index scores on motor characteristics, which in turn evaluates Parkinson's severity.
B. Speech Analysis utilizes audio input to accomplish binary classification based on voice characteristics.
C. A fusion method is developed to optimize the result based on different models' results.

ParkinAid was developed to facilitate at-home self-assessment for Parkinson's disease. By leveraging machine learning techniques, it predicts various Parkinson's index scores, offering a cost-effective and time-effective alternative to expensive hospital evaluations.

2. Theory and algorithm design

As described in Section 1, ParkinAid integrates motor and speech symptoms to provide a comprehensive assessment, enhancing diagnostic accuracy with high confidence. It deploys specialized models for different analysis:

1) Body motion analysis uses YoloV8 for feature extractions and uses Recurrent Neural Networks (RNN) with GRU, 1D-Convolutional Neural Networks (CNNs) and Logistic model (Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting Models, Logistic Regression (LR) and Dynamic Time Warping (DTW)) to analyze the movements, classifies PD stage, and predicts PD index scores.

2) Hand tremor analysis uses MediaPipe for feature extractions and NN-LSTM model, Graph Neural Networks (GNNs), and logistic models (NB, SVM, DT, RF, LR, KNN) to analyze the movement features and classifies PD stage.

3) Speech analysis uses NLP tool NLTK for semantic information extraction, ParselMouth for phonetic information extraction, OpenSMILE for prosodic information extraction, Kaldi for articulatory features extraction, and implements Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs) to accomplish binary classification.

2.1 Design

2.1.1 System Architecture

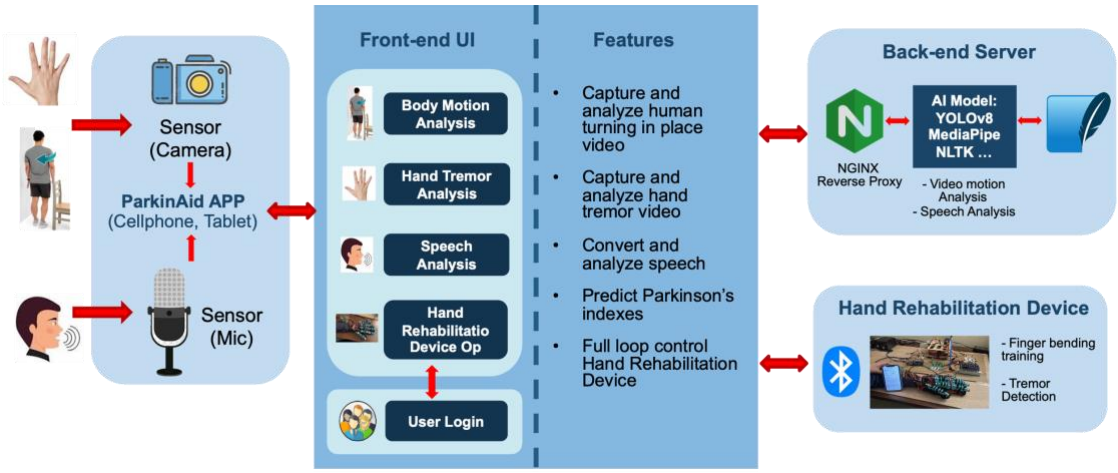


Figure 1 Architecture of ParkinAid

The working architecture of the ParkinAid is shown in Figure 1. The system includes data input and analysis. Audio and video input are obtained through camera and microphone. There's UI designed for each of the three functions described in section 1.3 respectively.

2.1.2 Data Analysis Workflow

The video and audio analysis aims to analyze the disorder of characteristics of movement and speech. The workflow of data analysis follows the same procedure as in Figure 2.

Feature Selection: Selecting the characteristics relevant to Parkinson's disease. The video characteristics are related to body anatomies. The audios analysis is to analyze speech semantics, articulation, phonation, and prosody. ParkinAid leverages all of the features to get the optimized result. This is so that ParkinAid doesn't miss out any possible chance of diagnosing. The selected features were all based on literatures review and UPDRS-III review [2][13]. The comprehensive summary of the features selected are shown in Table 1.

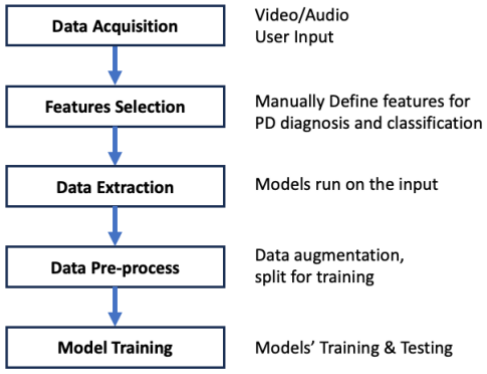


Figure 2: the workflow of data analysis for all models

Table 1: Motors and Features in UPDRS Part III

Motor Examination Items	Features in UPDRS	Features Selected
3.1 Speech	Volume, Rate, Prosody/modulation, Articulation clarity	Semantics, phonation prosody, articulation.
3.2 Facial Expression	Eye-blink frequency, Degree of expressiveness	
3.3 Rigidity	Resistance to passive movement, Range of Motion	Freezing of Gait Limb rigidity Freezing Events
3.4 Finger Tapping	Tapping speed, Amplitude of finger, Regularity, Decrement	
3.5 Hand Movements	Speed, Amplitude Hesitations or halts, Decrement	
3.6 Pronation-Supination Movements of Hands	Speed, Amplitude, Hesitations or halts, Decrement	
3.7 Toe Tapping	Tap speed, Amplitude, Hesitations or halts, Decrement	
3.8 Leg Agility	speed, amplitude, hesitations, halts and decrementing amplitude.	Freezing of Gait Limb rigidity
3.9 Arising from Chair	Speed, Amplitude, Hesitations or halts, Decrement	
3.10 Gait	stride amplitude, stride speed, height of foot lift, heel strike during walking, turning, and arm swing, but not freezing.	Freezing of Gait Limb rigidity

Motor Examination Items	Features in UPDRS	Features Selected
3.11 Freezing of Gait	start hesitation and stuttering movements especially when turning and reaching the end of the task.	Freezing of Gait Limb rigidity
3.12 Postural Stability	Number of steps needed for balance recovery, Center of gravity displacement, Force/acceleration of retropulsion	
3.13 Posture	flexion and side-to-side leaning	
3.14 Global Spontaneity of Movement / Body Bradykinesia	Overall movement frequency, Speed and amplitude of spontaneous movements, Reduced gesturing	
3.15 Postural Tremor of the Hands	the highest amplitude of rest tremor, onset/offset	Hand finger tremors
3.16 Kinetic Tremor of the Hands	Amplitude during finger-to-nose or finger-to-target movement, Frequency (Hz), Onset timing	
3.17 Rest Tremor Amplitude	the maximum amplitude of rest tremor for all four limbs and the lip/jaw	
3.18 Constancy of Rest Tremor	constancy of rest tremor for all four limbs and the lip/jaw	

*Note: Many examination items require specific action of the patients that are hard to be analyzed through computer methods.

Data pre-processing: for motion analysis, the raw output of features consists of key points coordinates representing tracked body/hand movements. To improve the robustness and generalizability of models trained on this data, several data augmentation techniques are applied, including noise injection, smoothing and interpolation, and generalization. For speech analysis, the augmentation of reverberation, speed perturbation, synthetic feature generation were deployed.

Model Training: The extracted features are used to train the ML models**;** actions include:

- Split data into training and testing sets.
- Train machine learning models.
- Perform hyperparameter tuning.
- Evaluate model performance (accuracy and reliability) of diagnosis and classification via metrics such as accuracy, precision, recall, and F1 score, etc.

A detailed workflow of ParkinAid data analysis is shown in Figure 3.

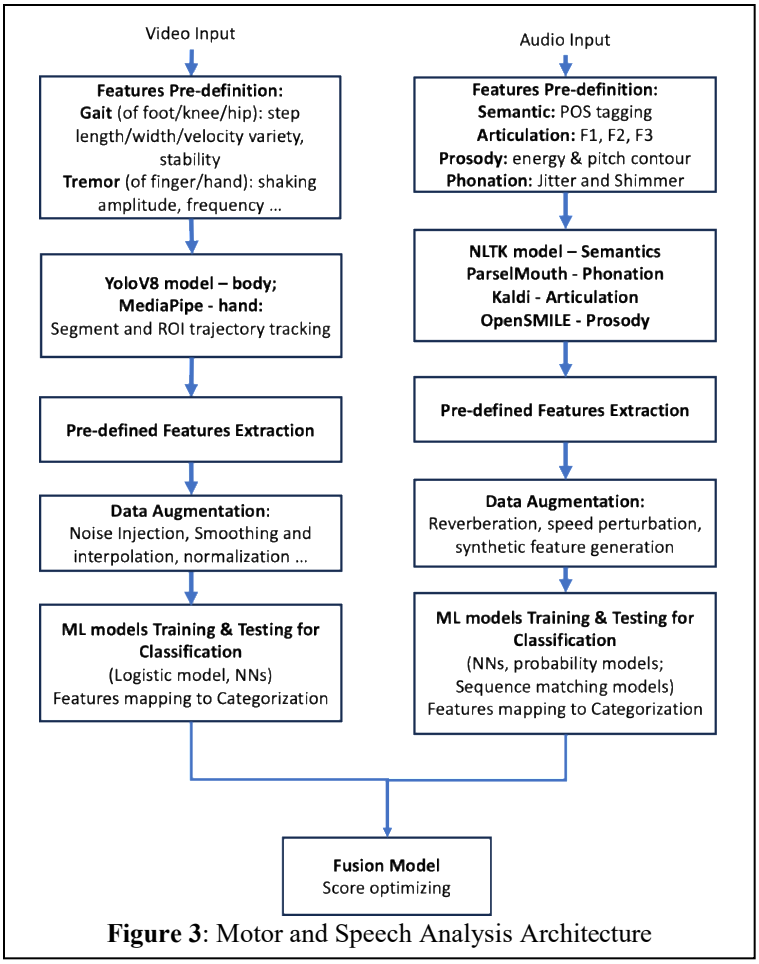


Figure 3: Motor and Speech Analysis Architecture

A fusion model integrates outputs from multiple individual models that perform different tasks in the ParkinAid system, such as binary classification and regression. The purpose of a fusion model is to take the diverse outputs into consideration and come up with a final decision score. The fusion model achieves this by the following:

$$Y = \max_{i \in \{1,2,\ldots,\eta\}} (\omega_i \cdot \frac{x_i}{S_i})$$

Where x_i is the output value of the model, S_i is the maximum score of scale in which x_i is evaluated in, and ω_i is the weight (F1-score) assigned to the model result. The maximum result across all model outputs will be the final fusion model result. The higher the result is, the higher

chance that the patient has Parkinson's. In the end, users will obtain individual results from all models ran and a result generated by the fusion model.

2.1.3 Input Data

The raw data of videos and audios were all obtained from publicly available dataset.

- The videos for body motion analysis come from figshare [15]. There are 80 videos covering patients with low to high Parkinson's severity. 70 are Patient with PD, 10 are normal persons without PD. The data are labeled with PD severity, including index: FoG Ratio, UPDRS-II, UPDRS-III, PIGD Score, Dyskinesia Score, MiniBestTest Score, TUG time, TUG dual-task time. The videos for hand tremor analysis come from shutterstock [16]. The individual video length is around 10-15 seconds each. 53 short videos of hand tremor are collected. The audios for the speech analysis come from UC Irvine Machine Learning Repository and figshare [17]. 30 voice samples between interviewers and PD patients are recorded. Each video is labeled with the PD patient's H&Y, UPDRS II-5, and UPDRS III-18 scores.

The data were randomly split into 80-20% for training and testing for both diagnosis and classification.

2.1.4 Feature Extraction

As described in section 2.1.1, the characteristics relevant to Parkinson's disease are related to body anatomies movement and speech features, identified through literatures reviews and guidance UPDRS-III review. After selecting these characteristics from Table 1, specific features are further refined for training and manual feature selection, which are shown in Table 2 and 3 below.

Motor characteristics detected are freeze of gait, hand tremor, and rigidity. Some features were obtained through calculation of raw data, which are extracted using YoloV8 and MediaPipe. The features data calculation methodology (i.e. Measurement Feature) are described. The detail of motor features and extractions tools used are shown in Table 2.

Table 2: Motor features, measurements, and extraction tools

Characteristics	**Features**	**Extraction Tool**	**Measurement Feature**
Freezing of Gait Limb rigidity	Angular Velocity	YoloV8	Change in rotation angle over time
	Smoothness of Rotation		Analyze acceleration/jerk
	Body Alignment/Posture		Track key points angles (head, shoulders, hips)
	Step Consistency		Variability in step timing, length
	Freezing Events		Detect sudden halts in motion
Hand finger tremors	Amplitude of tremor	MediaPipe	Displacement of key points overtime
	Finger tremor frequency		Rate of the oscillation
	Smoothness of movement		jerk

Acoustic characteristic detected are semantic information, speech phonation, articulation, and prosody. These features are extracted using different tools. The features data calculation methodology (i.e. Measurement Feature) are described. The detail of speech features and extractions tools used are shown in Table 3.

Table 3: Speech features, measurement and extraction tools

Characteristics	Features	Extraction Tool	Measurement Features
Semantics	Part-of-speech tagging	NLTK	Verb Rate, Common Noun Rate, Proper Noun Rate, Filler Word Rate
Phonation	Jitter (Variability in Pitch Frequency)	ParselMouth	Local Jitter, local absolute Jitter, rap Jitter, ppq5 Jitter, ddp Jitter

Characteristics	Features	Extraction Tool	Measurement Features
	Shimmer (Variability in amplitude)		Local Shimmer, local db Shimmer, apq3 Shimmer, aqpq5 Shimmer, apq11 Shimmer, dda Shimmer
	Harmonics-to-Noise Ratio		HNR
	Fundamental Frequency		Mean F0, Stdev F0
Articulation	Resonance frequencies of the speech tract	Kaldi	Formant Frequencies: F1, F2, F3
	Capture spectral properties of speech		MFCCs (Mel-Frequency Cepstral Coefficients)
Prosody	Pitch Contour	OpenSMILE	Mean F0, Stdev F0, Pitch Range, Pitch Slope
	Energy Contour		Mean Energy, Stdev Energy, Energy Range, Energy Slope
	Stress Patterns		Stress Energy Ratio, Inter-Stress Interval

All features, motion and acoustic, shown above are all considered in the model training process to mitigate issues that can happen in the manual feature selection as expertise is required for such process.

2.1.5 Evaluation Metrics

The performance of the trained ML and DL models on the test data was assessed using evaluation metrics: accuracy, precision, recall, F1-score, and mean squared error (MSE). Based on the outcome of a classification test, the number of true positives (TP) can be calculated as well as of true negatives (TN), false positives (FP) and false negatives (FN) [14]. Accuracy, which is one of the most used evaluation metrics, measures the proportion of correct predictions **Error! Reference source not found.** over the total number of evaluated instances:

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall represents the proportion of positive patterns that are correctly classified:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score measures the harmonic mean between recall and precision values:

$$F1\ score = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Mean squared error measures the average squared difference between the estimated values and the true value. This is used in ParkinAid as output values can be continuous predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$Where$:

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

Showing good result with these metrics indicates the success of the AI models.

2.2 Training

2.2.1 Model-Feature Mapping and Reasoning

This section details the mapping of specific features to machine learning models across body motion analysis, hand tremor analysis, and speech analysis, with a rationale for each assignment. The selection of models is based on their ability to address the spatial, temporal, and classification challenges inherent in the data. Various machine learning and deep learning algorithms were

deployed, evaluated, and compared to assess their effectiveness in handling these distinct analytical tasks. The models/algorithms deployed are shown in Table 4 and 5.

Table 4: Algorithms deployed for body motion/hand tremor analysis

Characteristics	Features	Measured Feature	Models Trained	Reasoning
Freeze of Gait Limb rigidity	Angular Velocity	Change in rotation angle over time	1D-CNN, RNN with GRU, DTW	Captures local patterns (CNN), temporal dependencies (GRU), and feature interactions (Boosting).
	Smoothness of Rotation	Analyze acceleration/jerk		Models sequence smoothness (DTW, GRU) with interpretable binary classification (LR).
	Body Alignment/Posture	Track key points angles (head, shoulders, hips)	Decision Tree, Random Forest, Naïve Bayes	Tree-based models handle discrete classifications; NB is lightweight for independent features.
	Step Consistency	Variability in step timing, length	RNN with GRU, Gradient Boosting Models, Logistic Regression	GRU detects temporal irregularities, boosting combines step variability features.
	Freezing Events	Detect sudden halts in motion	RNN with GRU, DTW, SVM	GRU detects temporal freezing patterns; DTW compares freezing sequences; SVM handles binary classification.
Hand finger tremors	Amplitude of tremor	Displacement of key pointss overtime	NN-LSTM, GNNs, Random Forest	LSTM captures temporal changes; GNN models spatial relationships; RF provides robust classification.
	Finger tremor frequency	Rate of the oscillation	NN-LSTM, GNNs, Decision Tree	LSTM learns periodic patterns; GNN exploits spatial relationships; KNN handles discrete frequency ranges.
	Smoothness of movement	jerk		LSTM tracks temporal smoothness; LR and DT offer interpretable smoothness classification.

Table 5: Algorithms deployed for speech analysis

Characteristics	Features	Measurement Features	Models Trained
Semantics	Part-of-speech tagging	Verb Rate, Common Noun Rate, Proper Noun Rate, Filler Word Rate	Hidden Markov Models (HMMs), Transformers (BERT)
Phonation	Jitter (Variability in Pitch Frequency)	Local Jitter, local absolute Jitter, rap Jitter, ppq5 Jitter, ddp Jitter	Gaussian Mixture Models (GMMs)
	Shimmer (Variability in amplitude)	Local Shimmer, local db Shimmer, apq3 Shimmer, aqpq5 Shimmer, apq11 Shimmer, dda Shimmer	
	Harmonics-to-Noise Ratio	HNR	
	Fundamental Frequency	Mean F0, Stdev F0	
Articulation	Resonance frequencies of the speech tract	Formant Frequencies : F1, F2, F3	Gaussian Mixture Models (GMMs), Convolutional Recurrent Neural Networks (CRNNs)
	Capture spectral properties of speech	MFCCs (Mel-Frequency Cepstral Coefficients)	
Prosody	Pitch Contour	Mean F0, Stdev F0, Pitch Range, Pitch Slope	Hidden Markov Models (HMMs), Temporal Convolutional Networks (TCNs)
	Energy Contour	Mean Energy, Stdev Energy, Energy Range, Energy Slope	
	Stress Patterns	Stress Energy Ratio, Inter-Stress Interval	

2.2.2 Body Motion Analysis Models Training

This section details the body motion analysis models training process, including the data augmentation and hyperparameters used. Input data were described in section 2.1.2. The models and their training parameters are shown in the Table 6.

2.2.2.1 Data Augmentation

When analyzing body motion videos where patients are turning in place, the raw output from YOLOv8 consists of key points coordinates representing tracked body movements. To improve the robustness and generalizability of models trained on this data, several data augmentation techniques are applied to the coordinate-based feature extraction result.

Noise injection is used to create more tremor-induced fluctuations by adding Gaussian noise and directional biases to key points trajectories, ensuring the model can handle erratic movement patterns. Smoothing and interpolation is used to refine the data by applying a moving average filter to reduce jitter and using linear interpolation to fill in missing frames, creating a more continuous representation of motion. Temporal scaling is used to alter movement speed to mimic Parkinsonian symptoms such as bradykinesia or compensatory motions, while spatial scaling adjusted the amplitude of movement to emphasize asymmetries, such as reduced arm swings.

To further standardize the dataset, coordinate normalization is done to align key points relative to the body's center of mass rather than absolute frame dimensions, ensuring consistency across different patients and video resolutions. Outlier is to eliminate extreme deviations caused by tremors or sensor errors using interquartile range (IQR) filtering, preserving meaningful motion data. Together, these techniques create a dataset that improves model robustness and generalization.

Figure 4 shows an example of the data augmentation process.

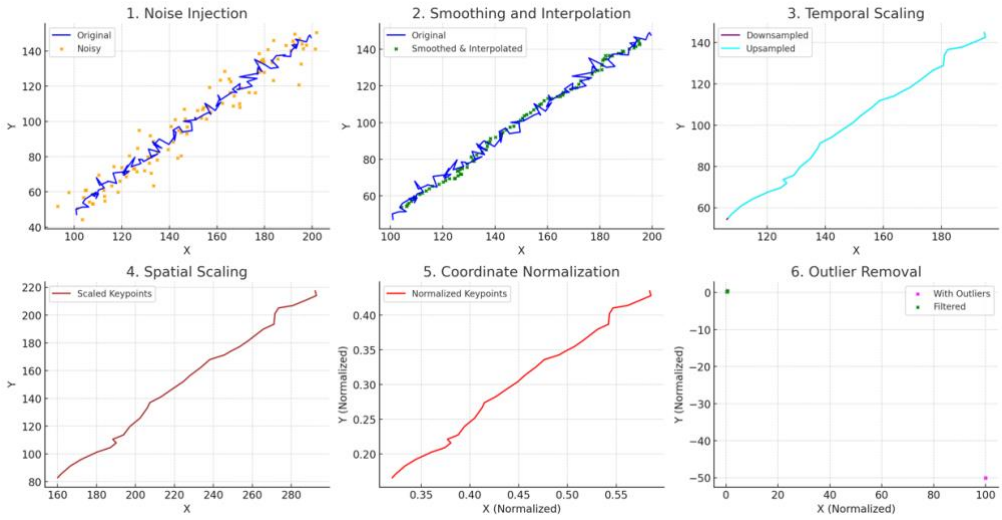


Figure 4: example of the data augmentation process

2.2.2.2 Hyperparameters

Below in Table 6 shows the hyperparameters and architecture configurations for all the models mentioned above. These are based on common practices for machine learning and deep learning models in the context of time-series or classification tasks like Parkinson's detection.

Table 6: hyperparameters for each model in body motion analysis

Model	Layers	Width	Activation Function	Learning Rate	Batch Size	Epochs	Optimizer	Dropout Rate	Loss Function	Turning	Notes
1D-CNN	6	64	ReLU	0.001	32	50	Adam	0.3	Cross-Entropy	Rotation Angles	Kernel size: 3; Max pooling used
RNN with GRU	4	128	Tanh	0.005	64	30	SGD	0.2	MSE	Sequence-Based	Sequence length: 50
DTW	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Pattern Matching	Dynamic alignment for rotation
Naïve Bayes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Log-Likelihood	Feature-Based	Gaussian features for body data
Gradient Boosting	N/A	N/A	N/A	0.1	N/A	N/A	N/A	N/A	Cross-Entropy	Angular Features	Depth: 5; Trees: 100

Model	Layers	Width	Activation Function	Learning Rate	Batch Size	Epochs	Optimizer	Dropout Rate	Loss Function	Turning	Notes
Logistic Regression	N/A	N/A	Sigmoid	0.01	64	N/A	SGD	N/A	Binary Cross-Entropy	Feature-Based	Regulization: L1 and L2 tested
SVM (RBF Kernel)	N/A	N/A	N/A	0.001	N/A	N/A	N/A	N/A	Hinge Loss	Rotation Angles	C: 1.0; Gamma: Scale
SVM (Linear Kernel)	N/A	N/A	N/A	0.001	N/A	N/A	N/A	N/A	Hinge Loss	Rotation Angles	C: 1.0; Linear separability assumed
SVM (Polynomial Kernel)	N/A	N/A	N/A	0.001	N/A	N/A	N/A	N/A	Hinge Loss	Rotation Angles	Degree: 3; C: 1.0

2.2.3 Hand Tremor Analysis Models Training

2.2.3.1 Data Augmentation

For detecting finger tremors, the raw data extracted through the MediaPipe model provides key points trajectories for hand and finger movements. The data augmentation process for finger tremor detection incorporates multiple techniques to enhance model robustness and generalization. High-frequency noise injection replicates involuntary micro-movements, ensuring the model can distinguish natural hand movement variability from pathological tremors. Voluntary hand movements, such as waving or grasping, are superimposed on tremor trajectories to help differentiate between normal and tremor-induced actions. Temporal distortions adjust the timing of movements by stretching or compressing

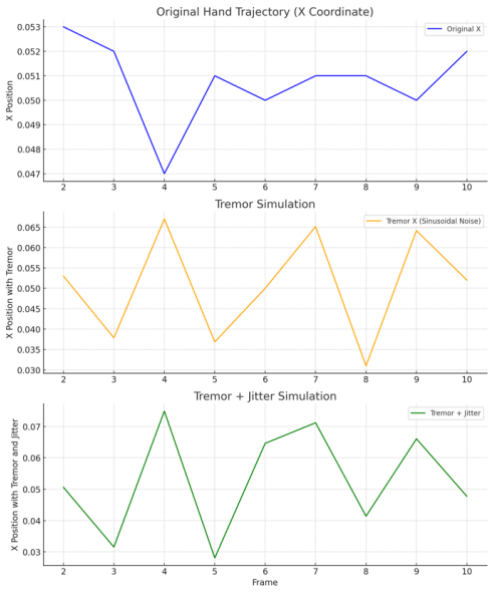


Figure 5: Tremor Data Augmentation Showcase

sections of the trajectory, mimicking irregular tremor intensity fluctuations. Additionally, synthetic tremor-free phases introduce periods of controlled motion, providing reference points for distinguishing tremor from normal hand activity. Together, these augmentations create a comprehensive dataset that accurately represents Parkinsonian tremor patterns in both static and dynamic hand conditions, improving the model's detection and classification capabilities.

2.2.3.2 Hyperparameters

Table 7: Hand tremor analysis models hyperparameters

Model	Layers	Width	Activation Function	Learning Rate	Batch Size	Epochs	Optimizer	Dropout Rate	Loss Function	Variation/kernel	Notes
NN-LSTM (Bi-Directional)	2	128	ReLU	0.005	64	50	Adam	0.4	Mean Squared Error	Bi-Directional	Sequence length: 100
NN-LSTM (Vanilla)	3	256	Tanh	0.001	32	30	SGD	0.3	Binary Cross-Entropy / MSE	Vanilla LSTM	Used for feature extraction
GNN (Fully Connected)	4	128	LeakyReLU	0.0005	32	100	AdamW	0.5	Cross-Entropy / MSE	Fully Connected Graph	Captures spatial movement relationships
GNN (DAG)	3	64	Sigmoid	0.001	64	50	SGD	0.2	Mean Absolute Error / MSE	Directed Acyclic Graph	Focused on directed graph movements
Random Forest (Entropy)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Entropy / MSE	Depth: 5; Trees: 50	Balanced for speed and accuracy
Random Forest (Gini)	N/A	N/A	N/A	N/A	N/A	N/A	SGD	N/A	Gini Impurity / MSE	Depth: 10; Trees: 100	Focuses on fine-grained decision boundaries
SVM (RBF Kernel)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Gini Impurity / MSE	Maximum Depth: 5	Simplified tree for explainability
Decision Tree (Simplified)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Cross - Entropy/ MSE	Maximum Depth: 10	Handles higher complexity
Decision Tree (Complex)	N/A	N/A	N/A	0.001	N/A	N/A	N/A	N/A	Hinge Loss	Rotation Angles	Degree: 3; C: 1.0

2.2.4 Speech Analysis Models Training

2.2.4.1 Data Augmentation

To improve model robustness against environmental distortions, reverberation is introduced by applying convolution filters with room impulse responses (RIRs), allowing the system to handle variations in microphone placement and echo effects. Time masking is implemented by muting short segments (50-200 ms) of audio or MFCC features to simulate missing data, ensuring models can extract meaningful patterns despite interruptions. Additionally, speed perturbation modifies the playback speed of speech samples by ±5% without altering pitch, enabling the model to generalize across different speaking rates and speaker styles.

Further augmentation techniques enhance speaker variability and class representation. Vocal tract length perturbation (VTLP) alters the spectral frequency axis, simulating speaker-dependent anatomical differences, which strengthens the model's ability to generalize across diverse vocal tract structures. Synthetic feature generation using Variational Autoencoders (VAEs) expands the dataset by generating realistic feature vectors for underrepresented speech patterns, mitigating class imbalances. Collectively, these augmentations refine the training data, enabling more robust speech analysis models capable of handling real-world inconsistencies in Parkinson's speech assessment.

2.2.4.2 Hyperparameters

Table 8: Speech analysis models hyperparameters

Model	Layers	Width	Activation Function	Learning Rate	Batch Size	Epochs	Optimizer	Dropout Rate	Loss Function	Variation/kernel	Notes
HMM (Basic Transition)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Negative Log-Likelihood / MSE	Constant Transition	Assumes stationary transitions
HMM (Dynamic Transition)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Negative Log-Likelihood / MSE	Adaptive Transition	Dynamically adjusts transitions over time
BERT (Base Fine-Tuned on MFCCs)	12	768	GELU	0.0001	16	10	AdamW	0.1	Cross-Entropy / MSE	Base Pretrained	Fine-tuned on MFCCs, pitch, and shimmer
BERT (Large Fine-Tuned on Stress)	24	1024	GELU	0.00005	8	20	AdamW	0.2	Cross-Entropy / MSE	Large Pretrained	Fine-tuned on inter-stress intervals and energy
GMM (Full Covariance)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Log-Likelihood	Full Covariance	Models feature interdependencies explicitly
CRNN (Stacked GRUs)	3 Conv + 2 GRU	128	Tanh	0.0005	64	30	SGD	0.3	Cross - Entropy / MSE	Stacked GRU	Handles longer sequential dependencies
TCN (Residual Blocks)	6	256	LeakyReLU	0.0005	32	100	AdamW	0.2	MSE	Residual Blocks	Deeper network with residual connections

3. Results

The testing phase aims to validate the accuracy, robustness, and generalizability of the ParkinAid system using the testing dataset. Each component—body motion analysis, hand tremor detection, and speech evaluation—underwent rigorous testing. This section outlines the individual model effectiveness, performance summary, and discussions of the results.

3.1 Individual Model Effectiveness

The results for each individual model of body motion analysis are shown below in Table 9:

Table 9: results for each individual model of body motion analysis

Feature	Model	Accuracy	Precision	Recall	F1 Score	MSE
Smoothness of Rotation	1D-CNN	79.4%	76.9%	80.0%	78.4%	0.029
	RNN with GRU	87.5%	85.7%	90.0%	87.8%	0.022
	DTW	67.5%	65.0%	68.4%	66.7%	0.041
Body Alignment Posture	Decision Tree	75.0%	75.0%	75.0%	75.0%	N/A
	Random Forest	82.5%	85.0%	81.0%	82.9%	N/A
	Naïve Bayes	65.0%	66.7%	60.0%	63.2%	N/A
Step Consistency	RNN with GRU	88.0%	88.0%	88.0%	88.0%	0.020
	Gradient Boosting Models	79.0%	79.6%	78.0%	78.8%	0.028
	Logistic Regression	55.0%	55.6%	50.0%	52.7%	0.065
Freezing Events	RNN with GRU	90.0%	90.0%	90.0%	90.0%	0.018
	SVM (RBF Kernel)	75.0%	77.8%	70.0%	73.7%	N/A
	SVM (Linear Kernel)	70.8%	72.0%	69.0%	70.4%	N/A
	SVM (Polynomial Kernel)	72.5%	74.0%	71.0%	72.5%	N/A

The results for each model of hand tremor analysis are shown below in Table 10:

Table 10: results for each model of hand tremor analysis

Feature	Model	Accuracy	Precision	Recall	F1 Score	MSE
Amplitude of Tremor	NN-LSTM (Bi-Directional)	91.2%	90.5%	92.0%	91.2%	0.015
	GNN (Fully Connected)	88.7%	87.9%	89.4%	88.6%	0.021
	Random Forest (Entropy)	79.8%	80.2%	79.5%	79.8%	N/A
	Random Forest (Gini)	81.0%	80.5%	81.5%	81.0%	N/A
Finger Tremor Frequency	NN-LSTM (Vanilla)	90.3%	89.6%	91.0%	90.3%	0.018
	GNN (DAG)	87.2%	86.5%	88.0%	87.2%	0.025
	SVM (RBF Kernel)	74.5%	75.2%	73.9%	74.5%	N/A
	Decision Tree (Simplified)	62.3%	63.5%	60.8%	62.1%	N/A
	Decision Tree (Complex)	70.5%	72.0%	69.5%	70.7%	N/A

The results for each model of speech analysis are shown below in Table 11:

Table 11: results for each model of speech analysis

Feature	Model	Accuracy	Precision	Recall	F1 Score	MSE
Semantics	HMM (Basic Transition)	81.0%	80.0%	82.0%	81.0%	N/A
	BERT (Base Fine-Tuned on MFCCs)	91.5%	91.0%	92.0%	91.5%	0.012
	BERT (Large Fine-Tuned on Stress)	93.1%	92.8%	93.5%	93.1%	0.010
Phonation	GMM (Full Covariance)	78.5%	77.9%	79.0%	78.4%	N/A
Articulation	GMM (Full Covariance)	76.3%	75.5%	77.0%	76.2%	N/A
	CRNN (Stacked GRUs)	89.5%	89.0%	90.0%	89.5%	0.019
Prosody	HMM (Dynamic Transition)	80.2%	79.0%	81.5%	80.2%	N/A
	Temporal Convolutional Networks	87.6%	87.0%	88.0%	87.5%	0.022

3.2 Performance

The best results of each feature for all models shown above are summarized as below in Table 12, which are the most optimal result.

Table 12: best results of each feature for all models

Analysis Category	Metric	Accuracy	Precision	Recall	F1-Score	MSE
Body Motion Analysis	Mean	75.94%	76.25%	74.65%	75.39%	0.0319
	Max	90.00%	90.00%	90.00%	90.00%	0.0650
	Min	55.00%	55.60%	50.00%	52.70%	0.0180
Hand Tremor Analysis	Mean	80.61%	80.66%	80.62%	80.60%	0.0198
	Max	91.20%	90.50%	92.00%	91.20%	0.0250
	Min	62.30%	63.50%	60.80%	62.10%	0.0150
Vocal Analysis	Mean	84.71%	84.02%	85.37%	84.68%	0.0158
	Max	93.10%	92.80%	93.50%	93.10%	0.0220
	Min	76.30%	75.50%	77.00%	76.20%	0.0100

In speech analysis, the mean accuracy was 84.71%, with the highest accuracy (93.1%) recorded by BERT (Large Fine-Tuned on Stress for Semantics) and the lowest accuracy (76.3%) recorded by GMM (Full Covariance for Articulation). The models in this category had a mean precision of 84.02%, with values ranging from 75.5% (GMM for Articulation) to 92.8% (BERT - Large Fine-Tuned on Stress). Recall values ranged between 77.0% (GMM for Articulation) and 93.5% (BERT - Large Fine-Tuned on Stress). The mean MSE was 0.0158, with the highest MSE (0.022) observed in TCN for Prosody and the lowest MSE (0.010) in BERT - Large Fine-Tuned on Stress.

In hand tremor analysis, models had an average accuracy of 80.61%, with accuracy values spanning from 62.3% (Decision Tree - Simplified for Finger Tremor Frequency) to 91.2% (NN-LSTM Bi-Directional for Amplitude of Tremor). The mean precision was 80.66%, ranging from 63.5% (Decision Tree - Simplified for Finger Tremor Frequency) to 90.5% (NN-LSTM Bi-Directional for Amplitude of Tremor). Recall values varied between 60.8% (Decision Tree - Simplified for Finger Tremor Frequency) and 92.0% (NN-LSTM Bi-Directional for Amplitude of Tremor). The mean MSE in this category was 0.0198, with the highest MSE (0.025) observed in GNN - DAG for Finger Tremor

Frequency, while the lowest MSE (0.015) was recorded in NN-LSTM Bi-Directional for Amplitude of Tremor.

In body motion analysis, the mean accuracy was 75.94%, with models achieving accuracy scores ranging from 55.0% (Logistic Regression for Step Consistency) to 90.0% (RNN with GRU for Freezing Events). The mean precision was 76.25%, with the lowest precision at 55.6% (Logistic Regression for Step Consistency) and the highest at 90.0% (RNN with GRU for Freezing Events). Recall values were between 50.0% (Logistic Regression for Step Consistency) and 90.0% (RNN with GRU for Freezing Events). The mean MSE was 0.0319, with the highest MSE (0.065) in Logistic Regression for Step Consistency, and the lowest MSE (0.018) in RNN with GRU for Freezing Events.

3.3 Discussion

The results obtained from the models demonstrate that different classifiers excel in different aspects of Parkinson's Disease (PD) detection. While some classifiers achieve high recognition rates in hand tremor analysis, others outperform in body motion analysis or speech feature classification. The overall performance indicates that deep learning-based models (NN-LSTM, CRNN, GNNs) outperform simpler models (Decision Trees, Naïve Bayes, SVM with Linear Kernels), but classical methods still show significant effectiveness in specific cases.

A. Hand Tremor Analysis

For hand tremor amplitude and frequency prediction, the NN-LSTM (Bi-Directional) model achieved the highest accuracy (91.2%), precision (90.5%), and recall (92.0%), demonstrating its ability to capture temporal dependencies in tremor motion. Similarly, GNNs (DAG and Fully Connected) performed strongly by modeling spatial relationships in key points displacements over time, making them effective for movement-based analysis.

However, Decision Tree (Simplified and Complex) models showed significantly lower accuracy (~62-70%), highlighting their limitations in capturing temporal dependencies and continuous variations. This suggests that for tremor-related analysis, models capable of handling sequential or spatially dependent data are necessary to achieve robust classification performance.

Interestingly, Random Forest (Entropy and Gini) showed moderate performance (~79-81%), indicating that while ensemble methods improve decision-making over simple trees, they still struggle compared to deep learning models in handling fine-grained tremor variations.

B. Body Motion Analysis

For body motion analysis, including step consistency and freezing events, RNN with GRU consistently outperformed other models with 90.0% accuracy, precision, and recall due to its ability to process sequential step irregularities and motion transitions. Gradient Boosting and Random Forests performed well in body alignment posture classification, where the data is more structured, achieving accuracy scores above 79%.

However, SVM (RBF, Linear, and Polynomial Kernels) struggled with freezing event classification, achieving an average accuracy of only 70-75%. The non-sequential nature of SVM models makes them less suitable for tasks that require temporal dependencies to identify sudden stops in motion. Interestingly, DTW (Dynamic Time Warping) performed moderately (67.5%), reflecting its ability to measure motion pattern variations but its limitations in higher-dimensional classifications.

This suggests that deep learning-based sequential models (RNNs, NN-LSTMs) dominate when continuous movement tracking is required.

C. Speech Analysis

Speech analysis shows a clear distinction between generative probabilistic models (GMMs), deep learning models (CRNNs, TCNs), and sequence-based models (HMMs, BERT). The best-performing

models for phonation and articulation features were CRNNs (Stacked GRUs) and BERT (Large Fine-Tuned on Stress), which achieved accuracy scores of 89.5% and 93.1%, respectively. This suggests that deep learning models are particularly effective at capturing high-dimensional spectral and prosodic relationships.

However, for prosody and phonation, the HMM (Dynamic Transition) model and GMM (Full Covariance) performed moderately well (~78-80%), indicating that these probabilistic models can effectively model short-term variations in pitch and shimmer but lack the ability to learn complex temporal dependencies.

This highlights the importance of feature selection: in speech analysis, models perform significantly better when key features (e.g., MFCCs, jitter, shimmer, formants) are carefully chosen. In future work, feature selection methods such as SFFS (Sequential Forward Floating Selection) may be able to improve classifier efficiency, reducing redundancy in high-dimensional voice data.

D. Feature Integration Challenges and Considerations

A key takeaway from this study is that simply combining features from different modalities does not necessarily improve performance. Though, in some cases, integrating features does help boosting the model accuracy: in speech analysis, integrating phonation, articulation, and prosody features into a multi-class system was tried (though not shown) led to an improvement in classification accuracy for the SVM model up to 91.2%. The results overall still indicate that individual feature system provide the highest accuracy. This suggests that feature fusion must be approached carefully, as combining suboptimal features can dilute the effectiveness of stronger predictors.

E. Fusion Model

Recall that the fusion model takes independent model outputs and uses their respective F1 score as the weights, since the independent models are proven accurate and generally resistant to noise in this

section, the fusion model performs fine. However, further improvements in result integration are essential to enhance performance.

F. Final Observations

The study overall demonstrates that deep learning models (NN-LSTM, CRNNs, BERT) consistently outperform traditional models (SVM, Decision Tree, Naïve Bayes) in time-dependent tasks, but simpler models can still be useful for structured data classification. Furthermore, ensemble models like Random Forest and Gradient Boosting serve as strong middle-ground approaches, particularly when dealing with structured posture and articulation analysis.

3.4 Limitations

The current approach to Parkinson's diagnosis faces challenges. In the model training process, there's limited control over the dataset collected from the public sources given the extreme limited amount of data that can be found regarding diseases. When using ParkinAid, user's poor video and audio quality affects analysis, leading to results to be off. Computationally, deep learning models demand significant resources, limiting real-time feasibility. Addressing these issues will improve clinical applicability.

4. Conclusion

The ParkinAid system has demonstrated high accuracy, robustness, and reliability across its motion and speech analysis modules. ParkinAid is also a first attempt to standardize diagnosis under the same scale: it sets up a platform where future developers can keep adding on to the existing methods.

Future improvements will focus on:

- Expanding the dataset to include more diverse patient profiles
- Developing enhanced diagnosis features
- Developing advanced techniques to intelligently integrate voice, motion, and tremor analysis.

With these advancements, ParkinAid will continue evolving into an accessible, accurate, and proactive tool—ultimately improving patient outcomes and quality of life for those affected by Parkinson's Disease.

References:

[1] Han, L.. An Auxiliary Rehabilitation Device for Parkinson's Patients With Finger Muscle Tremors and Stiffness https://abstracts.societyforscience.org/Home/FullAbstract?ISEFYears=0%2C&Category=Any%20Category&AllAbstracts=True&FairCountry=Any%20Country&FairState=WI&ProjectId=25134

[2] Christopher G. Goetz, et al., "Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results", Movement Disorders Vol. 23, No. 15, 2008, pp. 2129–2170.

[3] Mei, J., Desrosiers, C., Frasnelli J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. Frontiers in Againg Neuroscience. Volume 13, 2021. https://doi.org/10.3389/fnagi.2021.633752.

[4] Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis.J. Neurol. Neurosurg. Psychiatry79, 368–376. doi: 10.1136/jnnp.2007.131045

[5] Stewart, A.F.; William, J.W. Parkinson's Disease: Diagnosis & Clinical Management, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2008

[6] ALI, M. R., SEN, T., LI, Q., LANGEVIN, R., MYERS, T., DORSEY, E. R., SHARMA, S., AND HOQUE, E. Analyzing head pose in remotely collected videos of people with parkinson's disease. ACM Trans. Comput. Healthcare 2, 4 (sep 2021)

[7] LANGEVIN, R., ALI, M. R., SEN, T., SNYDER, C., MYERS, T., DORSEY, E. R., AND HOQUE, M. E. The park framework for automated analysis of parkinson's disease characteristics. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 2 (jun 2019).

[8] SIBLEY, K., GIRGES, C., HOQUE, E., AND FOLTYNIE, T. Video-based analyses of parkinson's disease severity: A brief review. S83 – S93.

[9] WROGE, T. J., O¨ ZKANCA, Y., DEMIROGLU, C., SI, D., ATKINS, D. C., AND GHOMI, R. H. Parkinson's disease diagnosis using machine learning and voice. In Signal Processing in Medicine and Biology Symposium (2018), IEEE, pp. 1–7.

[10] JIN, B., QU, Y., ZHANG, L., AND GAO, Z. Diagnosing parkinson disease through facial expression recognition: video analysis. Journal of medical Internet research 22, 7 (2020), e18697.

[11] Ferreira MIASN, Barbieri FA, Moreno VC, Penedo T, Tavares JMRS. Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters. Gait Posture. 2022 Oct;98:49-55. doi: 10.1016/j.gaitpost.2022.08.014. Epub 2022 Aug 20. PMID: 36049418.

[12] Islam, M., Lee, S., Abdelkader, A. PARK: Parkinson's Analysis with Remote Kinetic-tasks. https://doi.org/10.48550/arXiv.2311.12654

[13] Ferreira MIASN, Barbieri FA, Moreno VC, Penedo T, Tavares JMRS. Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters. Gait Posture. 2022 Oct;98:49-55. doi: 10.1016/j.gaitpost.2022.08.014. Epub 2022 Aug 20. PMID: 36049418.

[14] A. Dumortier, E. Beckjord, S. Shiffman, E. Sejdic, Classifying smoking urges via machine learning, Comput. Methods Programs Biomed. 137 (2016) 203–213. doi:10.1016/j.cmpb.2016.09.016.

[15] https://figshare.com/articles/dataset/A_public_dataset_of_video_acceleration_and_angular_velocity_in_individuals_with_Parkinson_s_disease_during_the_turning-in-place_task/14984667?file=31324702

[16] https://www.shutterstock.com/video/clip-1106381543-hand-tremor-parkinson-patient-psychiatric-clinic-man

[17] https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127