Video-Based Assessment of Motor Skills in Parkinson's Disease Using Motion Representations from MotionGPT

Vasiliki Georgali
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
georgaliv@iti.gr

Sofia Balula Dias
CIPER, Faculdade de Motricidade Humana
Universidade de Lisboa
Lisbon, Portugal
sbalula@fmh.ulisboa.pt

Beatriz Alves
Faculdade de Motricidade Humana
Universidade de Lisboa
Lisbon, Portugal
ft.beatriz.alves@gmail.com

Sevasti Bostantjopoulou
Third Neurological Clinic
Papanikolaou Hospital
Thessaloniki, Greece
bostkamb@otenet.gr

Zoe Katsarou
Third Neurological Clinic
Papanikolaou Hospital
Thessaloniki, Greece
katsarouzoe@gmail.com

Nikos Grammalidis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
ngramm@iti.gr

Kosmas Dimitropoulos
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
dimitrop@iti.gr

***Abstract*—Accurate assessment of motor skills is critical in patients with neurodegenerative diseases like Parkinson's Disease. Motor skill assessment is an essential tool for early diagnosis, tracking disease progression, and guiding effective treatment and management. Since assessment often requires continuous monitoring, advances in computer vision and machine learning provide an opportunity to streamline this process by enabling automated and non-invasive evaluation methods. In this work, we present a novel framework for automated motor skill assessment that integrates pose estimation, parametric body modeling and motion representation learning. Unlike conventional video-based assessment methods, our framework does not rely on RGB video input but operates solely on skeletal landmarks and leverages a pretrained motion encoder to learn rich latent motion representations. The proposed pipeline begins with video recordings of patients performing standardized motor tasks, from which skeletal landmarks are estimated using MediaPipe. To enable compatibility with upstream motion models, we propose a deep learning-based mechanism to map MediaPipe landmarks to SMPL pose parameters, aligning with MotionGPT's pretraining on SMPL-based motion representations. The resulting motion sequences are processed by the pretrained MotionGPT motion encoder to obtain latent representations that capture spatiotemporal dynamics. These latent representations are then used to train time-series classifiers, including LSTM-based and Transformer-based architectures, for predicting clinically relevant motor skill scores. We evaluate our approach against baselines based on raw landmarks and handcrafted features, using both deep learning and traditional machine learning algorithms. The proposed approach outperforms these baselines, demonstrating the effectiveness of leveraging features derived from pretrained motion models for motor skill assessment.**

Keywords— motor skill assessment, Parkinson's Disease, motion representation, MotionGPT

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder that begins years before diagnosis can be made, affects multiple neuroanatomical areas, arises from a combination of genetic and environmental factors, and manifests with a broad range of both motor and non-motor symptoms. The most prevalent motor impairments, include tremor, slowness in movement (bradykinesia), stiffness and postural instability [1].With the progression of the disease, these symptoms significantly affect the patient's quality of life and their ability to perform daily activities. Even though there is currently no cure for PD, its symptoms can be partially managed with medication. Therefore, continuous monitoring of motor symptoms is essential for effective disease management. Clinical assessment of motor skills is typically performed using standardized rating scales such as the Unified Parkinson's Disease Rating Scale (UPDRS) and its updated version, the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [2], [3] .The MDS-UDPRS test includes self-assessment components in the form of self-administrated questionnaires to evaluate how a patient perceives their own daily functioning, however a substantial portion of the test is performed by medical professionals and relies on the clinician's expert assessment of subtle impairments or motor changes through observation of a patient performing predefined motor tasks.

Despite the fact that the MDS-UDPRS is a well-established and widely adopted method for assessment of PD symptoms, it has some notable limitations. Since most items on the test are scored by a medical professional, the assessment depends on the clinician's expertise and interpretation and is therefore inherently subjective, introducing inter-and intra-rater variability in scoring [4], [5]. Moreover, since evaluations are typically conducted during clinical visits, this type of assessment offers only a snapshot of the patient's conditions rather than a continuous monitoring of symptom progression. Another limitation of MDS-UDPRS is that it is a time-consuming process, especially for patients residing at a distance from healthcare centers. As a result, there is increasing interest in developing objective and continuous monitoring tools that can complement or enhance traditional clinical assessment methods.

A substantial body of research has explored the use of technology-driven approaches for objective and continuous motor assessment. Early research focuses on the use of

wearable devices or smartphones, collecting and processing sensor data, such as accelerometer and gyroscope signals to capture motion impairments during daily activities [6], [7]. Such systems have demonstrated effectiveness in quantifying tremor, gait abnormalities and bradykinesia [8], [9]. More recently, vision-based approaches have emerged as a promising alternative, leveraging advances in the field of computer vision and machine learning to enable analysis of human motion from video recordings. Markerless pose estimation frameworks, such as OpenPose [10] and MediaPipe [11], enable real-time extraction of 3D skeletal keypoints from monocular video. These tools have been widely adopted in the analysis of motor skills across a range of medical conditions, including PD, eliminating the need for specialized hardware or markers [12], [13].

Multiple studies have demonstrated that pose-based handcrafted features can be effectively used for automated assessment of PD. In [14], an early vision - based framework is presented, which combines pose estimation and deep learning techniques to predict clinical scores related to dyskinesia. In a similar manner, [15] proposes a video-based system for quantitative gait analysis, where kinematic features are extracted from pose sequences and used to assess disease severity. In [16], MDS-UPDRS scores are estimated using handcrafted features from MediaPipe sequences and classical machine learning models. Both works demonstrate that video-based methods can approximate clinical evaluations, reducing the necessity for in-person assessments.

Other works have focused on fine motor assessment using video-based hand tracking [17]. In [18], utilizes MediaPipe to analyze hand movements and estimate tremor characteristics before and after medication in PD patients, achieving accurate quantification of tremor frequency and amplitude. Similarly, [19] investigated the use of multiple hand pose estimation models for assessing bradykinesia in remote settings, highlighting both the potential but also the challenges of deploying such assessment systems in real-time.

With the rapid advancements in the field of neural networks, an increasing number of studies move beyond handcrafted features, employing deep learning models to capture spatio-temporal dynamics in motion data [20]. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) architectures, are often being used to model short/long-term dependencies in motor tasks, while Transformer-based architectures have gained attention due to their ability to capture long-range temporal dynamics. For example, [21] developed a two-channel model to learn spatio-temporal patterns in gait data of PD patients, which combines LSTMs and Convolutional Neural Networks (CNN). The proposed system showed better performance compared to traditional machine learning algorithms. A Transformer-based architecture is introduced in [22] that leverages both temporal and spatial characteristics of 1D gait signals to differentiate between healthy individuals and PD patients.

The emergence of large-scale motion models, such as MotionGPT [23], has introduced powerful mechanisms for learning high-level spatiotemporal dynamics from motion data, achieving notable performance in motion related tasks, such as action recognition, motion prediction and motion synthesis. However, the adoption of such models in clinical contexts - and particularly in PD assessment - remains limited. Existing works [24], [25] that incorporate motion-language models primarily exploit its language modeling capabilities to generate textual assessments. Moreover, these approaches do not address self-assessment or scalable, patient-facing monitoring scenarios.

In this context, we propose a unified framework that combines video-based pose estimation with pretrained motion encoding to extract latent spatio-temporal features from PD patient motor assessments. By aligning the input pose data with the requirements of motion language models, our approach moves beyond conventional landmark-based pipelines relying on handcrafted features, and provides a more expressive basis for motor assessment. In summary, the key contributions of this study are:

- Motion embeddings derived from a pretrained motion-language model are utilized to capture rich spatiotemporal dynamics, showcasing their effectiveness in automated motor skill assessment.
- A deep learning-based regression framework is proposed to transform MediaPipe landmarks into SMPL-H parameters [26], [27], a widely adopted parametric human body representation, allowing the use of pretrained models build on SMPL-based motion data.
- The proposed framework is validated on video recordings of real patients performing standardized motor tasks, demonstrating its applicability in clinically relevant settings.

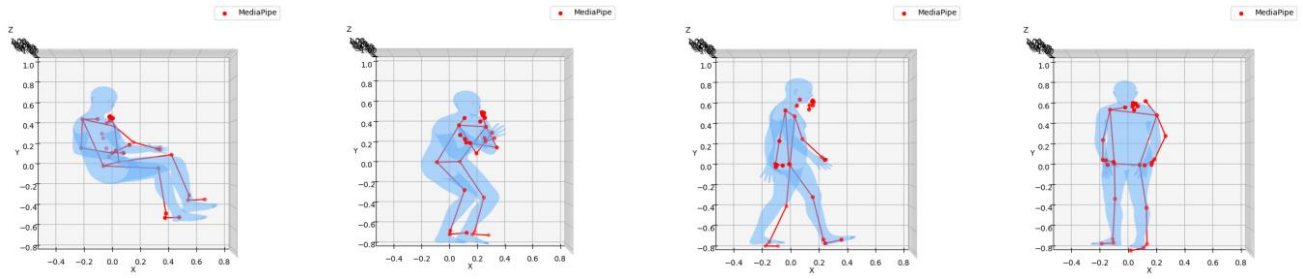
II. METHODOLOGY

A. Data Collection Protocol

A video-based analysis method was developed to evaluate motor function in individuals with Parkinson's disease (PD), with particular attention to balance and postural control. The designed approach enables automated assessment of selected motor tests from the MDS-UDPRS. Patients were instructed to perform four motor tests: "item 3.8" (leg agility), "item 3.9" (arising from chair), "item 3.10" (gait) and "item 3.13" (posture). While the MDS-UDPRS protocol was followed as closely as possible, minor adjustments were necessary, since the tests were designed for self-administration, meaning patients might need to record themselves without assistance from a medical professional. The selected motor items are performed consecutively and recorded in a single continuous sequence.

The four chosen motor tests of the proposed Comprehensive Motor Function Test (CMFT) are described as follows:

- **Leg agility (item 3.8)**: The patient is seated in a straight-backed chair with arms, with their feet comfortably on the floor. The patient is then instructed to raise and stomp their foot on the ground 10 times as high and as fast as possible. Each leg is tested separately.
- **Arising from chair (item 3.9)**: From the seated position, the patient is instructed to cross their arms across their chest and then stand up. If the patient is not successful, they should try to repeat this attempt up to a maximum of two more times. If still unsuccessful, the patient is allowed to push off the chair using their arms.



3.8 (Leg agility) *3.9 (Arising from chair)* *3.10 (Gait)* *3.13 (Posture)*

Fig. 1. The proposed CMFT protocol featuring MDS-UDPRS items for motor skill assessment.
In red, the MediaPipe landmarks estimated from the MediaPipe Pose Landmarker and in light blue the 3D SMPL-H mesh representations.

- **Posture (item 3.13)**: After the patient stands up, they are asked to stand still for 5 seconds, so that their posture can be assessed.
- **Gait (item 3.10)**: After the posture assessment, the patient is instructed to take 3 steps away from the chair (3.10F), turn around and return to the chair, taking 3 steps again (3.10B). The original MDS-UDPRS test requires the patient to walk at least 10m, but due to the limited field of view of the camera, this test was adapted.

This protocol provides a reliable framework for standardized and consistent video recordings suitable for automated analysis, while preserving compatibility with clinically validated motor assessment methods. Fig. 1 demonstrates the MDS-UPDRS items in the proposed CMFT protocol. To ensure compliance with patient privacy rights, for visualization purposes we use 3D SMPL-H mesh representations instead of actual recorded footage of the patients. The meshes have been created using the SMPL-H pose parameters estimated from the MediaPipe-to-SMPL Pose Regressor network described in the following section.

B. Methods

The designed pipeline begins with extracting the 3D pose from the video recordings using the MediaPipe pose landmarker. The estimated skeleton is augmented with extra spine joints and aligned to the SMPL coordinate system and then used to estimate SMPL pose parameters in a frame-wise manner. Instead of processing frames independently through the SMPL model, the estimated pose parameters for the entire sequence are temporally concatenated and converted to a 3D joint representation in a single batch. The latter are converted into the HumanML3D [28] representation and encoded into a motion sequence using the pretrained MotionGPT Vector Quantized Variational Autoencoder (VQ-VAE). Finally, the encoded sequence is processed by a time-series classifier to produce the predicted label. These steps are explained in more detail in the following sections. A schematic overview of the proposed pipeline is shown in Fig. 2.

a) Landmark extraction and preprocessing

The recorded videos of the PD patients are first processed using Google's MediaPipe Pose Landmarker which employs the BlazePose algorithm to extract 33 3D landmarks (joints) per frame, either in world coordinates or in normalized pixel coordinates. For the purpose of this work, we use the world-coordinate landmarks. The use of MediaPipe enables the extraction of skeletal keypoints directly from video streams, ensuring that only motion data are processed and stored. This is particularly important for protecting patient privacy, as it eliminates the need to store or process visual data, in contrast to pipelines that directly map video input to parametric body models [29], [30], [31]. Additionally, MediaPipe's lightweight design, makes it suitable for scalable and interactive applications, including remote and continuous self-administrated motor assessments. These landmarks provide an initial skeletal representation of the recorded human motion. However, unlike SMPL, which provides detailed spine joints as part of its parametric model, BlazePose does not estimate landmarks for the spine region. To bridge the gap between the two representations and drawing inspiration from [32], 4 additional spine landmarks are estimated by interpolating positions between the hip midpoint and the shoulders midpoint. Subsequently, to ensure consistency in orientation and axis conventions between the two representations, the augmented skeleton is transformed appropriately to align with the SMPL coordinate system.

b) SMPL-H Pose & Joint Estimation

An SMPL-H representation is required to ensure compatibility with the pretrained motion encoder described in the following section. Therefore, the final augmented skeletons are flattened and passed through a neural network to predict the SMPL-H pose parameters θ_t. SMPL-H typically uses 15 additional joints for each hand, but since hand articulation is not relevant to our task, we set these parameters to zero and focus on predicting just the pose parameters for the pelvis and the 21 body joints, resulting in $\theta_t \in \mathbb{R}^{66}$. The first three values of the predicted vector correspond to the axis angle representation of the pelvis (root) joint and describe the global orientation of the body, while the rest correspond to the axis angle representations of the remaining 21 body joints and describe how each body joint is rotated relative to its parent in the kinematic chain. The MediaPipe-to-SMPL Pose Regressor network consists of four sequential blocks with decreasing hidden dimensions ($512 \rightarrow 512 \rightarrow 256 \rightarrow 128$), each composed of one linear layer followed by batch normalization and ReLU activation. Dropout regularization with a rate of 0.2 is applied to the outputs of the first two blocks to mitigate overfitting. A final linear layer produces the predicted SMPL-H pose parameters. Inference is performed on a frame-by-frame basis, without enforcing any temporal

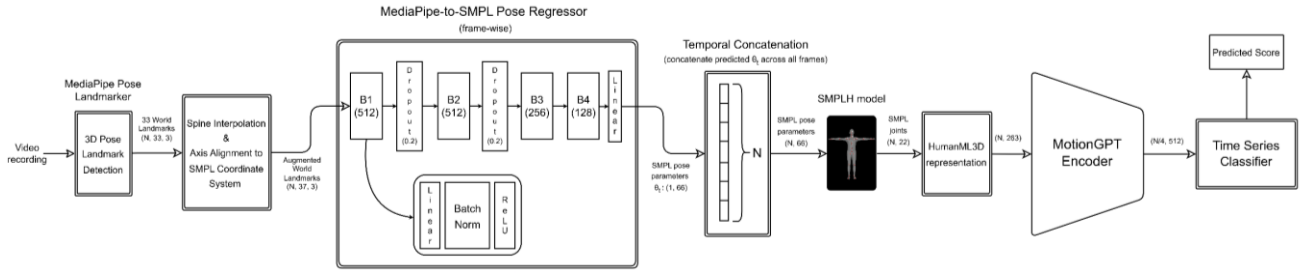


Fig. 2. Overview of the proposed pipeline for video-based motor assessment.

constraints. The predictions are then concatenated along the time dimension to form a sequence $\Theta = (\theta_1, \theta_2, \ldots, \theta_N)$, where N is the number of frames in the recorded video. This sequence is then passed through the SMPL-H forward model to acquire the 3D positions of the SMPL-H joints resulting in a final sequence of shape $J \in \mathbb{R}^{N \times 22 \times 3}$.

c) Motion Representation and Encoding

To leverage the capabilities of the pretrained MotionGPT encoder, we adopt the HumanML3D motion representation which was used during its training. This conversion maps the SMPL-H motion data to the encoder's expected feature space ($M \in \mathbb{R}^{N \times 263}$), enabling effective transfer without retraining. Furthermore, to align with the expected temporal resolution, motion sequences are resampled to 20 fps - while preserving the initial frame count through interpolation. These HumanML3D motion representations are then processed by the motion encoder Ɛ, part of MotionGPT's Vector Quantized Variational Autoencoder to yield latent temporal representations, also incorporating a temporal down sampling rate l of 4. While Ɛ employs a vector quantization module to map the latent motion representations into discrete codebook tokens, we omit this component and instead use the encoder's continuous latent representations directly, thereby avoiding the information loss introduced by quantization.

d) Time Series Classification

The encoded motion sequence is finally passed to a Time Series Classifier, which is responsible for predicting the MDS-UDPRS score for each motor test. To evaluate different temporal modeling strategies, two architectures are considered:

- An LSTM-based classifier, which captures temporal dependencies through recurrent processing and hidden state propagation and
- A Transformer-based classifier, which models long-range dependencies using self-attention mechanisms and aggregates temporal information via mean pooling.

In both cases, the sequence is reduced to a fixed-length representation, which is then passed through a fully connected layer to yield the predicted class probabilities.

III. Experimental Results

A. Datasets

a) AMASS dataset

The AMASS dataset [33] is used to train and evaluate the Mediapipe-to-SMPL pose regressor. The AMASS dataset is a large database of human motion, unifying different optical marker-based motion capture datasets by representing them within a common framework and parametrization (SMPL). We use the available SMPL-H parameters, which include additional parameters for hand-articulation. Since hand articulation is not relevant to our task, we set all hand pose parameters to zero and only use the body pose parameters. In order to create MediaPipe-to-SMPL mapped pairs, we first use the provided SMPL parameters to render videos of the available motion sequences. We then process the rendered videos using the MediaPipe pose landmarker, to get the corresponding 3D MediaPipe world landmarks.

For training and validation, we use subsets ACCAD, BMLhandball, BMLmovi, BMLrub, CMU, DanceDB, DFaust, EKUT, EyeJapanDataset, Human4D, HumanEva, KIT, Mosh, PosePrior, SFU, SOMA, SSM, TCDHands and Transitions. All subsets are split subject-wise so that every subject and its recorded sequences appears exclusively in one of the splits. We also test the performance of the model on the held-out subsets HDM05 and TotalCapture.

b) AI – PROGNOSIS Papanikolaou CMFT dataset

As part of the AI-PROGNOSIS initiative (https://ai-prognosis.eu), a cohort of 54 PD patients were recruited in the Papanikolaou General Hospital of Thessaloniki (Greece) and in the Northern Greece Parkinson's Disease Association to voluntarily participate in this study. All patients provided written informed consent in order to participate in the study. Participants were recorded using a smartphone placed in a fixed position (e.g., mounted on a tripod). The camera was positioned in such a way as to provide a clear field of view of a chair and sufficient surrounding space for the patient to walk three steps away from the chair, turn and walk back while remaining within the frame. This setup allowed all selected motor tasks of the CMFT protocol to be performed and recorded in a single video. A total of 85 videos were initially collected using this procedure, with some participants being recorded more than once. However, 35 of these recordings were discarded for various reasons, (e.g., patient moved outside the field of view, patient experiencing freezing of gait (FoG), errors in pose estimation by MediaPipe), yielding a final dataset that consists of 50 videos by 39 patients. Two independent neurologists from the Papanikolaou General Hospital of Thessaloniki were asked to assess the recordings and provide MDS-UDPRS scores

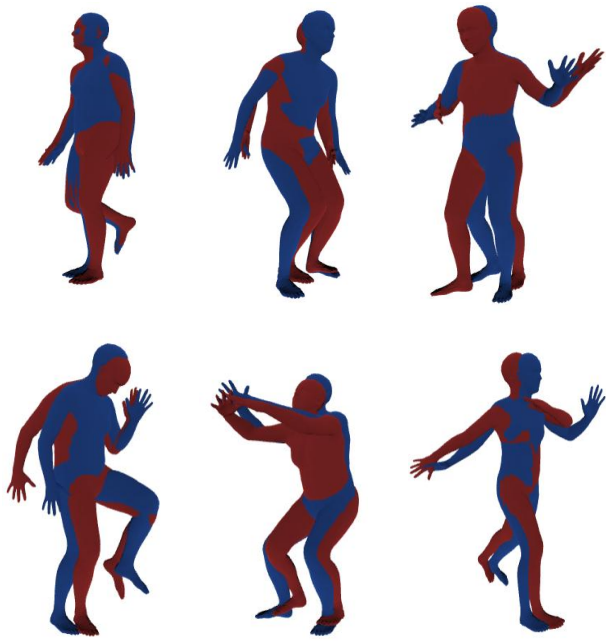


Fig. 3. 3D SMPL-H mesh reconstructions from the held-out test set. Ground truth meshes are shown in red and predicted meshes are shown in blue.

for each item of the CMFT protocol, ranging between 0 and 3 for all items.

B. *Mediapipe-to-SMPL Pose Regressor*

The Mediapipe-to-SMPL pose regression network was trained using a composite objective to balance joint-level accuracy and pose plausibility. The loss function is defined as follows:

$$L = L_{pose} + \lambda L_{joints} \quad (1)$$

where L_{pose} is defined as the geodesic loss between predicted vs ground truth pose parameters and L_{joints} is defined as the mean squared error (MSE) between predicted vs ground truth SMPL joint positions. Also, the weight λ is set to 20. The model was trained for 30 epochs, with early stopping applied after 5 epochs without improvement. Optimization is performed using AdamW optimizer with learning rate l_r=0.001 and weight_decay=0.0003. To quantitatively evaluate reconstruction performance, we also report the Mean Per Joint Position Error (MPJPE) achieved on the AMASS dataset: 59.15 mm on the training set and 65.56 mm on the validation set.

Fig. 3 illustrates qualitative comparisons between predicted vs ground truth SMPL-H poses sampled from the held-out test set using the corresponding 3D mesh representations, which convey human pose and body shape more comprehensively than skeletal representations. Overall, the predicted SMPL-H poses align well with the underlying motion, capturing both global orientation and body joint rotations, however accuracy decreases for distal joints, with Mean Per Joint Position Error (MPJPE) increasing for joints that are further from the body core, such as wrists and ankles. Fig. 4 presents visual

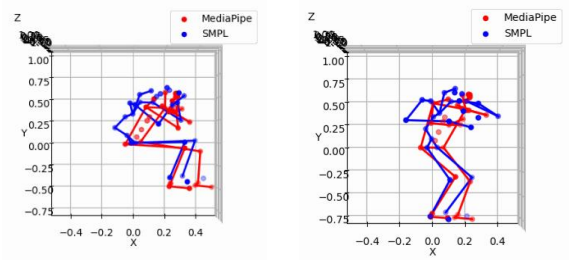


Fig. 4. Visual comparison between input Mediapipe skeleton (red) and predicted SMPL skeleton (blue) for a patient performing motor item 3.9 (arising from chair).

comparisons between the input MediaPipe skeletons and the predicted SMPL reconstructed skeleton for a patient performing motor item 3.9.

C. *Motor Item Score Prediction*

For the downstream task of score prediction for each motor item, both LSTM-based and Transformer-based models are trained using a cross-entropy loss function. For comparison, we also trained traditional ML models to predict the motor item scores based on hand-crafted feature vectors computed from the time series of key angles computed for each motor item (e.g. knee angle for item 3.8). Five classifiers were employed, i.e., K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forest (RF), principal component analysis (PCA) in combination with KNN, and supervised PCA (sPCA) [34] in combination with KNN.

Since scores for the motor items range from 0 to 3, we formulate the task as a four-class classification problem, where each score corresponds to one distinct class. To evaluate the performance of the different architectures, we use stratified group k-fold cross validation. The folds are made by preserving the class distribution of the entire dataset, while also utilizing a subject-wise partitioning strategy. The latter ensures that, within each fold, a single subject's recordings appear either in the training or the test set.

We adopt the weighted F1-score as the primary evaluation metric, which considers both precision and recall for each class, while accounting for class imbalance. As seen from Table I, the LSTM- and Transformer-based architectures are the ones that achieve the best F1 scores for all motor items. More specifically, the LSTM-based architecture achieves the best F1 scores in all motor items except item 3.9 (arising from chair), in which case the Transformer-based architecture achieves the best performance, with the LSTM-based model achieving the second-best performance among all the classifiers. In all cases, except motor item 3.8, both the best and second-best performances are accomplished by the time-series classifiers, with the LSTM-based architecture usually outperforming the Transformer-based approach, likely due to the relative limited size of the dataset. In the case of motor item 3.8, the second-best performance is achieved by a classical machine learning architecture (PCA + KNN), with the top-performing architecture (LSTM) only marginally better.

TABLE I. AVERAGE WEIGHTED F1 SCORES FOR ASSESSMENT OF DIFFERENT MOTOR ITEMS, USING DIFFERNET CLASSIFIERS AND GROUP K-FOLD CROSS VALIDATION.

Classifier	Motor Item				
	3.8 (mean)	***3.9***	***3.13***	***3.10F***	***3.10B***
KNN	0.459	0.649	0.622	0.326	0.378
SVM	0.416	0.562	0.639	0.356	0.353
RF	0.468	0.575	0.614	0.351	0.421
PCA+KNN	**0.485**	0.617	0.589	0.323	0.384
sPCA+KNN	0.410	0.581	0.612	0.332	0.364
MediaPipe + LSTM	0.303	0.443	0.429	0.328	0.348
Angle timeseries + LSTM	0.247	0.485	0.409	0.325	0.283
MotionGPT + LSTM	**0.489**	**0.714**	**0.771**	**0.604**	**0.673**
MotionGPT + Transformer	0.441	**0.745**	**0.702**	**0.454**	**0.589**

* Bold and underlined values indicate the best performance

**Bold values indicate the second-best performance.

For item 3.8, across all trained models, F1 scores range from 0.4097 (sPCA +KNN) to 0.4888 (MotionGPT + LSTM) forming a rather narrow band. This suggests that the more complex architectures do not necessarily deliver a clear and consistent improvement, possibly because the motion under assessment (leg agility) is relatively subtle. For the remaining motor items, the time-series classifiers substantially outperform the classical machine learning models. For motor item 3.9 the Transformer-based classifier achieves an F1 score of 0.7447 while the best performing classical machine learning model (KNN) achieves a score of 0.6490. This is 14.75% improvement. Comparing the best F1 scores achieved by the time-series classifiers and the best F1 scores achieved by the classical machine learning models, the following improvements are achieved: 20.61% for motor item 3.13, 69.40% for motor item 3.10F and 59.87 for motor item 3.10B.

To assess the contribution of the intermediate MotionGPT representation, we additionally trained and evaluated two alternative time-series classifiers: one operating directly on the raw MediaPipe landmarks and another using joint-angle time series relevant to each motor item, computed from the original MediaPipe landmarks. Across all motor-items, the classifiers trained on raw MediaPipe landmarks or angle time series, perform substantially worse than the approaches trained on the latent temporal representations.

Overall, the results demonstrate that in almost all cases the latent temporal MotionGPT representations provide a more informative and robust representation than both handcrafted features and raw MediaPipe landmarks.

IV. CONCLUSIONS AND DISCUSSION

In this work, we presented a novel framework for automated motor skill assessment in patients with Parkinson's Disease. The proposed framework integrates pose estimation, parametric body modeling and learned motion representations. By mapping MediaPipe landmarks to SMPL-H parameters, the pipeline enables the use of pretrained motion models to extract informative spatio-temporal representations for downstream clinical assessment tasks. Experimental results demonstrate that leveraging motion embeddings trained on SMPL-based datasets, significantly improves performance in motor skill assessment, in most items in the proposed CMFT protocol. Both LSTM- and Transformer-based classifiers consistently outperform models trained on raw landmarks or hand-crafted features.

Despite the strong overall performance of the pipeline, one limitation should be noted: the MediaPipe-to-SMPL pose regression module is less accurate for distal joints (e.g. hands and feet). This is particularly relevant in clinical settings, where subtle distal movements are important and inaccuracies in these body regions may introduce uncertainty in the motion representation. This limitation is evident in the score prediction for motor item 3.8 (leg agility) where the motion under assessment is more subtle. Another notable shortcoming of this study was the exclusion of 35 out of 85 recordings due to tracking failures. This is a substantial percentage of the total number of recordings, especially when thinking about scaling up the analysis. This issue is particularly relevant on intermediate to later stages of the disease, because FoG is triggered by gait initiation and changes of direction, and the CMFT only allows for 3 steps to be taken before turning so it can exacerbate FoG. This limitation also affects the applicability of the framework in real-world self-assessment settings, where tracking failures, occlusions and freezing episodes are expected to occur.

Future work includes improving the modeling of distal joints and evaluating the framework on larger and more diverse patient cohorts. Further research should also investigate more robust tracking approaches and modified assessment protocols to improve reliability and scalability.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union under grant agreement no. 101080581 (AI-PROGNOSIS: Artificial intelligence-based Parkinson's disease risk assessment and prognosis).

REFERENCES

[1] L. V. Kalia and A. E. Lang, "Parkinson's disease," Aug. 29, 2015, *Lancet Publishing Group*. doi: 10.1016/S0140-6736(14)61393-3.

[2] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, no. 1, pp. 41–47, Jan. 2007, doi: 10.1002/mds.21198.

[3] C. 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, pp. 2129–2170, Nov. 2008, doi: 10.1002/mds.22340.

[4] L. J. W. Evers, J. H. Krijthe, M. J. Meinders, B. R. Bloem, and T. M. Heskes, "Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS," *Movement Disorders*, vol. 34, no. 10, pp. 1480–1487, Oct. 2019, doi: 10.1002/mds.27790.

[5] L. M. D. Luiz, I. A. Marques, J. P. Folador, and A. O. Andrade, "Intra and inter-rater remote assessment of bradykinesia in

Parkinson's disease," *Neurologia*, vol. 39, no. 4, pp. 345–352, May 2024, doi: 10.1016/j.nrl.2021.08.005.

[6] R. Matias, V. Paixão, R. Bouça, and J. J. Ferreira, "A perspective on wearable sensor measurements and data science for Parkinson's disease," *Front. Neurol.*, vol. 8, no. DEC, Dec. 2017, doi: 10.3389/fneur.2017.00677.

[7] G. AlMahadin, A. Lotfi, E. Zysk, F. L. Siena, M. M. Carthy, and P. Breedon, "Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective," *BMC Neurol.*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12883-020-01996-7.

[8] H. Li, M. Zecca, and J. Huang, "Evaluating the utility of wearable sensors for the early diagnosis of parkinson disease: systematic review," 2025, *JMIR Publications Inc.* doi: 10.2196/69422.

[9] T. Chatzis, N. Grammalidis, and K. Dimitropoulos, "A deep learning approach for parkinsonian tremor assessment using wearables," in *2025 IEEE International Conference on E-health Networking, Application and Services, Healthcom 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/HealthCom60686.2025.11342654.

[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," May 2019, [Online]. Available: http://arxiv.org/abs/1812.08008

[11] MediaPipe. 2026. Pose landmark detection guide (April 2026). Retrieved May 11, 2026 from https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker.

[12] G. Sarapata *et al.*, "Video-based activity recognition for automated motor assessment of Parkinson's disease," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 10, pp. 5032–5041, Oct. 2023, doi: 10.1109/JBHI.2023.3298530.

[13] L. di Biase, P. M. Pecoraro, and F. Bugamelli, "AI video analysis in Parkinson's disease: A systematic review of the most accurate computer vision tools for diagnosis, symptom monitoring, and therapy management," Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s25206373.

[14] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *J. Neuroeng. Rehabil.*, vol. 15, no. 1, p. 97, Dec. 2018, doi: 10.1186/s12984-018-0446-z.

[15] K. Abe, K. I. Tabei, K. Matsuura, K. Kobayashi, and T. Ohkubo, "Relationship between the results of arm swing data from the OpenPose-based gait analysis system and MDS-UPDRS scores," *IEEE Access*, vol. 10, pp. 118992–119000, 2022, doi: 10.1109/ACCESS.2022.3220767.

[16] A. Stergioulas *et al.*, "Assessing motor skills in Parkinson's Disease using smartphone-based video analysis and machine learning," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2024, pp. 562–568. doi: 10.1145/3652037.3663945.

[17] G. Amprimo, G. Masi, G. Olmo, and C. Ferraris, "Deep learning for hand tracking in Parkinson's Disease video-based assessment: Current and future perspectives," Aug. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.artmed.2024.102914.

[18] G. Güney *et al.*, "Video-based hand movement analysis of Parkinson patients before and after medication using high-frame-rate videos and MediaPipe," *Sensors (Basel)*, vol. 22, no. 20, Oct. 2022, doi: 10.3390/s22207992.

[19] G. T. Acevedo Trebbau, A. Bandini, and D. L. Guarin, "Video-Based Hand Pose Estimation for Remote Assessment of Bradykinesia in Parkinson's Disease," in *Predictive Intelligence in Medicine. PRIME 2023*, I. Rekik, E. Adeli, S. H. Park, C. Cintas, and G. Zamzmi, Eds., Lecture Notes in Computer Science, vol. 14277, Cham, Switzerland: Springer, 2023, pp. —, doi: 10.1007/978-3-031-46005-0_21.

[20] S. B. Dias *et al.*, "Motion analysis on depth camera data to quantify Parkinson's Disease patients' motor status within the framework of I-Prognosis Personalized game suite," *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020, pp. 3264-3268, doi: 10.1109/ICIP40778.2020.9191017.

[21] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "A hybrid spatio-temporal model for detection and severity rating of Parkinson's disease from gait data," *Neurocomputing*, vol. 315, pp. 1–8, Nov. 2018, doi: 10.1016/j.neucom.2018.03.032.

[22] D. M. D. Nguyen, M. Miah, G.-A. Bilodeau, and W. Bouachir, "Transformers for 1D signals in Parkinson's Disease detection from gait," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.00423

[23] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "MotionGPT: Human motion as a foreign language." [Online]. Available: https://github.com/OpenMotionLab/MotionGPT

[24] D. Wang, C. Bobenrieth, and H. Seo, "AGIR: Assessing 3D gait impairment with reasoning based on LLMs," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.18141

[25] R. Yang, A. Kennedy, and R. J. Cotton, "BiomechGPT: Towards a biomechanically fluent multimodal foundation model for clinically relevant motion tasks," May 2025, [Online]. Available: http://arxiv.org/abs/2505.18465

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[27] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 245:1–245:17, Nov. 2017.

[28] C. Guo *et al.*, "Generating diverse and natural 3D human motions from text," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2022, pp. 5142–5151. doi: 10.1109/CVPR52688.2022.00509.

[29] J. Liu, N. Akhtar, and A. Mian, "Deep reconstruction of 3-D human poses from video," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 3, pp. 497–510, Jun. 2023, doi: 10.1109/TAI.2022.3164065.

[30] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," Jul. 2016, [Online]. Available: http://arxiv.org/abs/1607.08128

[31] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video Based Reconstruction of 3D People Models," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8387–8397

[32] M. Kim, T. Kim, and K. T. Lee, "3D digital human generation from a single image using generative AI with real-time motion synchronization," *Electronics (Switzerland)*, vol. 14, no. 4, Feb. 2025, doi: 10.3390/electronics14040777.

[33] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.03278

[34] B. Ghojogh and M. Crowley, "Unsupervised and supervised principal component analysis: Tutorial" Aug. 2022, [Online]. Available: http://arxiv.org/abs/1906.03148