

Motion2Meaning: A Clinician-Centered Framework for Contestable LLM in Parkinson’s Disease Gait Interpretation

Loc Phuc Truong Nguyen¹, Hung Thanh Do¹,
Hung Truong Thanh Nguyen², and Hung Cao²

¹ Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany

² University of New Brunswick, Fredericton E3B 5A3, Canada

{loc.pt.nguyen,hung.t.do}@fau.de, {hung.ntt,hcao3}@unb.ca

Abstract. AI-assisted gait analysis holds promise for improving Parkinson’s Disease (PD) care, but current clinical dashboards lack transparency and offer no meaningful way for clinicians to interrogate or contest AI decisions. To address this issue, we present Motion2Meaning, a clinician-centered framework that advances Contestable AI through a tightly integrated interface designed for interpretability, oversight, and procedural recourse. Our approach leverages vertical Ground Reaction Force (vGRF) time-series data from wearable sensors as an objective biomarker of PD motor states. The system comprises three key components: a Gait Data Visualization Interface (GDVI), a one-dimensional Convolutional Neural Network (1D-CNN) that predicts Hoehn & Yahr severity stages, and a Contestable Interpretation Interface (CII) that combines our novel Cross-Modal Explanation Discrepancy (XMED) safeguard with a contestable Large Language Model (LLM). Our 1D-CNN achieves 89.0% F1-score on the public PhysioNet gait dataset. XMED successfully identifies model unreliability by detecting a five-fold increase in explanation discrepancies in incorrect predictions (7.45%) compared to correct ones (1.56%), while our LLM-powered interface enables clinicians to validate correct predictions and successfully contest a portion of the model’s errors. A human-centered evaluation of this contestable interface reveals a crucial trade-off between the LLM’s factual grounding and its readability and responsiveness to clinical feedback. This work demonstrates the feasibility of combining wearable sensor analysis with Explainable AI (XAI) and contestable LLMs to create a transparent, auditable system for PD gait interpretation that maintains clinical oversight while leveraging advanced AI capabilities. Our implementation is publicly available at: <https://github.com/hungdothanh/motion2meaning>.

Keywords: Parkinson’s disease care · Gait analysis · Human-centered contestable AI · Contestable large language model · Explainable AI

1 Introduction

The management of chronic neurodegenerative diseases is shifting from episodic evaluations to continuous monitoring with wearable sensors, which provide ob-

jective digital biomarkers for earlier intervention and individualized therapy [25,5,26]. Parkinson’s Disease (PD), a condition marked by progressive motor impairments [16], exemplifies this need. Standard clinical tools like the Unified Parkinson’s Disease Rating Scale (UPDRS) [27] are applied too intermittently and are vulnerable to observer and patient bias [6,2]. Consequently, they fail to capture daily motor fluctuations, leading to imprecise treatment, heightened fall risk, and reduced quality of life [26,17].

Although AI models can accurately quantify gait and predict disease severity [52,29], their clinical translation is stalled by a critical “last-mile problem.” Current dashboards present outputs like Hoehn & Yahr stages [12] as opaque, static scores, preventing clinicians from scrutinizing or overriding predictions that conflict with their expertise. This opacity undermines trust and the principles of evidence-based medicine. While Explainable AI (XAI) offers partial solutions like saliency maps [20,32,30], these are typically one-way communications that fail to support the dialogic nature of clinical reasoning [34,31]. The crucial gap is not merely a lack of transparency but the absence of mechanisms for procedural recourse, enabling clinicians to actively contest and amend AI-driven decisions.

To address this gap, we draw on the principles of Contestable AI (CAI). CAI extends beyond explanation by embedding structures for dialogue, challenge, and justification within system design [34,31]. This approach aligns with regulatory demands for human oversight (GDPR [48], EU AI Act [8]). A contestable system allows users not only to understand a decision but also to dispute it with domain expertise, ensuring that such challenges are recorded, processed, and capable of influencing the final outcome. This study seeks to apply these principles in a clinician-centered interface for PD care. The key contributions are as follows:

- We design and implement *Motion2Meaning*, a novel clinician-centered framework that unifies three core components: a deep learning (DL) diagnostic model, a dual-modality explainability module, and an LLM-driven interaction layer within a single human-in-the-loop interface.
- We implement a *1D-CNN architecture* that performs end-to-end classification of Hoehn & Yahr severity from raw vGRF time-series data. This model outputs a probability distribution over the four discrete severity stages.
- We introduce *Cross-Modal Explanation Discrepancy* (XMED), a novel XAI technique to automatically flag unreliable predictions. XMED operates on the principle that trustworthy predictions should have stable explanations across different methods. It quantifies the divergence between a gradient-based explanation (Grad-CAM [45]) and a backpropagation-based one (LRP [20]). A high divergence score signifies inconsistent model reasoning, which automatically flags the prediction for mandatory clinical review.
- We develop a *contestable interaction system* powered by a Large Language Model (LLM) that uses a structured “Contest & Justify” workflow. The LLM synthesizes the CNN’s prediction, XAI-identified salient features, and the clinician’s specific challenge to generate clinically-grounded textual justifications. These justifications form the basis for a transparent, evidence-based dialogue between the clinician and the AI.

2 Background and Related Work

This work is situated at the confluence of two research domains. We first review advances in AI for sensor-based PD gait analysis, where progress in predictive accuracy has often come at the cost of clinical interpretability. We then connect this gap to the broader evolution of human-centered AI in healthcare, which argues for moving beyond passive Explainable AI (XAI) toward the more interactive and legally robust principles of Contestable AI (CAI).

2.1 AI for Sensor-Based Parkinson’s Gait Analysis

For decades, gait analysis has been central to movement disorder research, but it was traditionally limited to specialized motion capture laboratories. Wearable Inertial Measurement Units (IMUs) have transformed this field by enabling continuous, high-resolution data collection in natural environments [43,24,40,38]. This is vital for PD care, where gait serves as a rich digital biomarker. Neurodegeneration of dopaminergic pathways in the basal ganglia disrupts movement automaticity, producing measurable deficits in stride length, speed, cadence, turning velocity, and asymmetry. IMUs are also well-suited to detect episodic phenomena such as Freezing of Gait (FOG) and medication-related fluctuations that are difficult to capture in clinic visits [43,35]. Early computational methods relied on handcrafted biomechanical features derived from statistical, spectral, and non-linear analyses, which were classified using models such as Support Vector Machines and Random Forests [37]. Although interpretable, these approaches were constrained by their dependence on expert-driven feature design and limited ability to capture complex pathological patterns. More recently, end-to-end DL has emerged, with one-dimensional Convolutional Neural Networks effective for local spatio-temporal motifs, and Recurrent Neural Networks or Transformers capturing long-range dependencies [7,28,39]. Despite SOTA performance, their opacity poses a major barrier to clinical adoption. This accuracy–interpretability trade-off erodes trust, as clinicians are reluctant to rely on opaque predictions that cannot be examined against their expertise [34,31]. The problem is compounded by dashboards that act only as data presenters, showing parameters or outputs without revealing model reasoning or enabling clinician input or correction [13,34].

2.2 From Explainable to Contestable AI in Healthcare

Bridging the gap between high-performance AI and clinical use requires human-centered socio-technical systems that are transparent, interpretable, and trustworthy. This effort began with XAI and is now advancing toward CAI, as illustrated in Figure 1 [31]. XAI seeks to make black-box predictions understandable, supporting trust and error detection [33,30,32]. In clinical gait analysis, XAI remains early, though methods from other medical domains provide guidance. Backpropagation-based approaches, such as Saliency Maps and LRP [20], highlight critical temporal regions of the input. CAM-based techniques, including Grad-CAM [45] and Grad-CAM++ [4], localize discriminative regions linked to

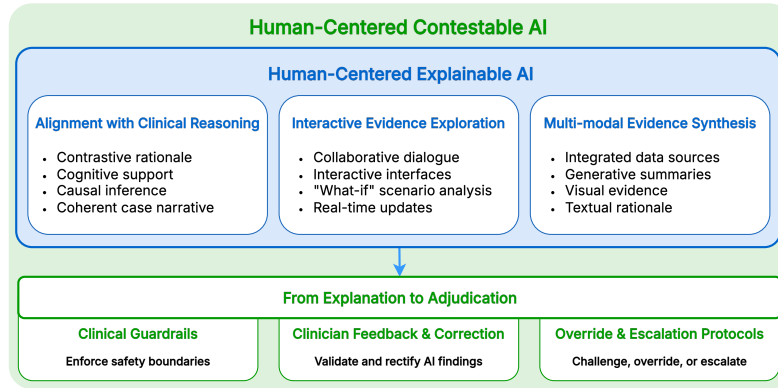


Fig. 1: Progression of human-centered XAI toward CAI.

predictions. Perturbation-based methods [41,36,49] identify influential regions by altering inputs and monitoring changes in output probabilities. More recent work emphasizes interactive explanations, allowing users to test counterfactuals or adjust inputs, and cognitively aligned formats, such as contrastive reasoning or natural language dialogue [34,31]. These developments recognize explanation as a social process aimed at shared understanding between humans and AI.

A central challenge in human-AI collaboration is achieving appropriate trust calibration [42,21], the process by which a user develops an accurate mental model of an AI’s capabilities to avoid both blind over-trust (automation bias) and reflexive dismissal (algorithm aversion). Conventional XAI, while providing transparency, may not suffice for this task. A compelling but incorrect explanation can actively impair calibration by creating a false sense of security [53]. A system designed for effective calibration must therefore go beyond one-way explanations and provide a mechanism for procedural recourse [11]. This is the principle of CAI: to create an essential feedback loop where clinicians can act on their calibrated judgments, transforming them from passive observers into active supervisors, enabling expert-driven recourse that forms the foundation of our Motion2Meaning framework, which aims to operationalize contestability in a real-world clinical setting.

3 Methodology

Figure 2 presents an overview of the Motion2Meaning framework, which integrates two core interfaces into an end-to-end system for gait interpretation.

The first component is the **Gait Data Visualization Interface (GDVI)**. It is an interactive web-based tool for exploring gait data from the PhysioNet dataset [10], as depicted in Figure 3. The interface presents raw vGRF signals in 10-second windows, with controls to toggle sensor channels and select time segments for detailed inspection. A complementary summary panel uses color-

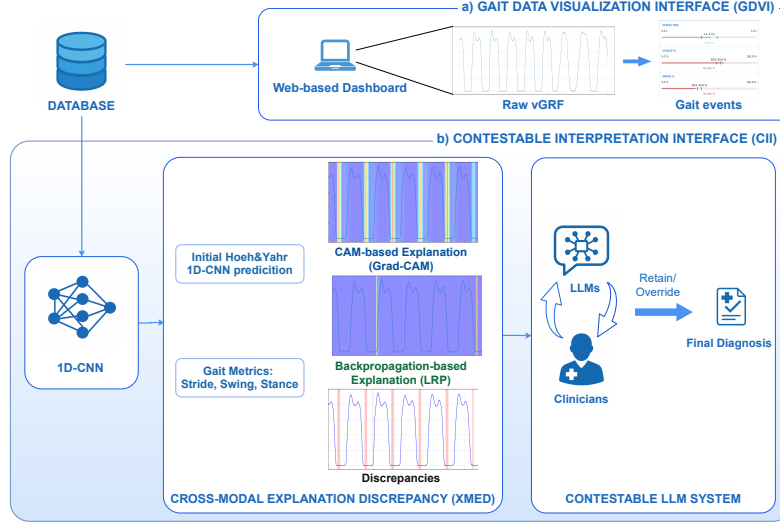


Fig. 2: The overview of Motion2Meaning framework: (a) Gait Data Visualization Interface (GDVI), and (b) Contestable Interpretation Interface (CII).

coded markers to highlight key temporal events, including Stride, Stance, and Swing Time. This design enables clinicians to rapidly compare gait patterns against normative data and visually identify potential anomalies.

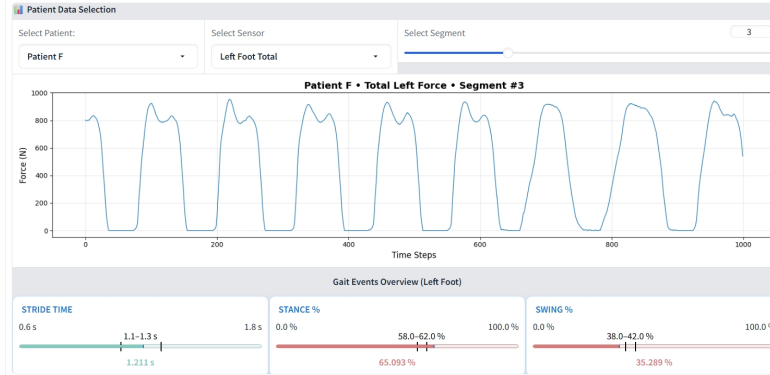


Fig. 3: The dashboard overview of Gait Data Visualization Interface (GDVI).

The second and also core component of the framework is the **Contestable Interpretation Interface (CII)**, the dashboard where the human-AI dialogue occurs. Its workflow is operationalized through three integrated technical pillars:

1. The workflow begins with the *1D-CNN*, our predictive engine designed following [1], which analyzes a patient’s gait data to generate an initial classification of the Hoehn & Yahr severity. This prediction is presented not as a final answer, but as a testable hypothesis for clinical review.
2. To audit the reliability of this hypothesis, the *Cross-Modal Explanation Discrepancy* (XMED) module performs an automated consistency check on the model’s reasoning. It leverages the fundamental differences between Grad-CAM [45], which identifies *where* in the signal the model focuses, and LRP [20], which attributes *what* specific data points were most influential. A significant divergence between these two reveals a critical failure mode where the model may correctly identify a clinically relevant temporal region but base its decision on a spurious artifact within it. This quantified "attention-attribution" gap provides a targeted alert for a structurally flawed reasoning process that a single explanation method would miss.
3. Finally, the *LLM-powered "Contest & Justify" workflow* enables procedural recourse. When a prediction is challenged, either due to an XMED alert or independent clinical judgment, the clinician registers dissent through a structured typology: *Factual Error* (contesting input data integrity), *Normative Conflict* (flagging contradiction with clinical knowledge), or *Reasoning Flaw* (challenging the XAI’s visual evidence). This formal contestation triggers the LLM to synthesize all available evidence and generate a new, contextualized justification. This exchange creates a transparent, collaborative negotiation that culminates in either acceptance or a clinician-driven override, with every step logged in an immutable record to ensure accountability.

4 Experiment and Results

To evaluate the effectiveness of our Motion2Meaning framework, we conducted a comprehensive, multi-stage investigation. Our evaluation was designed to answer three central questions: (1) *What is the baseline predictive performance of our core 1D-CNN model on PD severity classification?*; (2) *Can the XMED method effectively distinguish between reliable and unreliable model predictions?*; (3) *How effectively can LLMs leverage these discrepancy signals to validate correct predictions and contest erroneous ones in a simulated clinical workflow?*

4.1 Dataset and Experimental Setup

Our experiments were conducted using the public PhysioNet Gait in PD dataset, which contains vertical Ground Reaction Force (vGRF) signals from 93 individuals with PD and 73 healthy controls [10]. For our deep learning model, we preprocessed the data by segmenting the variable-length recordings into fixed, non-overlapping 1000-frame windows. To create a focused and interpretable attribution space for our XMED safeguard, we used a single, highly informative feature for our analysis: the "Total left force" signal.

The dataset was partitioned into training (70%), validation (15%), and test (15%) sets, using multiple random seeds to ensure robustness. We trained our model using a nested cross-validation strategy on the training data. A 5-fold outer loop assessed model generalization, while a 3-fold inner loop within each fold conducted a grid search to optimize hyperparameters. The final optimal configuration derived from this process is presented in Table 1.

4.2 Human-Centered Evaluation Metrics

To align with CAI and foreground the framework’s human-centered design, we evaluate Motion2Meaning using human-oriented metrics. We concentrate on LLM-generated textual explanations, which constitute the most direct interface between the AI component and clinicians. Our first two metrics are the Flesch readability tests [18,9], which estimate readability via sentence length and lexical complexity. **Flesch Reading Ease (FRE)** ranges from 1 to 100, with higher values indicating greater accessibility. The **Flesch-Kincaid Grade Level (FKGL)** estimates the U.S. school grade needed to comprehend a text. For clinician-oriented medical materials (e.g., clinical documentation used in diagnostic and care workflows), typical FRE scores are 50 to 70. These correspond to FKGL 8 to 12 and are appropriate for readers aged approximately 13 to 18 [51,3]. The corresponding formulas are given as:

$$\begin{aligned}\text{FRE} &= 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right), \\ \text{FKGL} &= -15.59 + 0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right).\end{aligned}$$

Clinical Grounding (CG) evaluates LLM hallucination by quantifying the verifiability of its explanations against available evidence. We compute it by first isolating all numerical values in the model’s generated text, and then determining the percentage of those numbers that match the figures provided in the input prompt and data. Given the multiset of numerical values extracted from the LLM’s generated explanation, V_E , and the multiset of all numerical values provided in the input prompt and data, V_I , CG is defined as:

$$\text{CG} = \frac{100}{|V_E|} \sum_{\nu \in V_E} \mathbb{I}(\nu \in V_I),$$

where $\mathbb{I}(\cdot)$ is the indicator function and the score is defined as 100 if $|V_E|=0$ (i.e., no numerical claims are made). A high score indicates the framework’s ability to reduce a clinician’s cognitive load and mitigate clinical risk.

Self-Correction Accuracy (SCA), measures the direct impact of our contestation system on fixing the baseline model’s mistakes. We compute it by first isolating the set of instances $\mathcal{D}_{\text{err}}$ that the baseline model initially

misclassified, and then calculating the percentage of these specific errors that the contestable LLMs system successfully overturns to the correct label. Given the final system’s prediction $\hat{y}_{\text{final}}(x_i)$ and the true label y_i for an instance $x_i \in \mathcal{D}_{\text{err}}$, SCA is formally defined as:

$$\text{SCA} = \frac{100}{|\mathcal{D}_{\text{err}}|} \sum_{x_i \in \mathcal{D}_{\text{err}}} \mathbb{I}(\hat{y}_{\text{final}}(x_i) = y_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function. A high score confirms the framework’s self-remediation capacity, enhancing user trust and clinician-AI collaboration.

4.3 Predictive Model Performance

Upon evaluation on the unseen test set, our model achieved a robust overall accuracy and weighted F1-score of 0.89, with a detailed breakdown of per-class performance provided in Table 2. The model demonstrates excellent performance in identifying the Healthy control group, achieving an F1-score of 0.91. Furthermore, it shows reliable discrimination between the clinically adjacent intermediate severity levels, with balanced F1-scores of 0.87 for Stage 2 and 0.89 for Stage 2.5. The primary challenge was observed in the most advanced category, Stage 3, which recorded a slightly lower recall of 0.84. This reduced sensitivity is likely attributable to the significant class imbalance, as this category contains only 83 samples, which may limit the model’s ability to learn its full intra-class variability. These results confirm the model’s overall robustness but also highlight a clear direction for future refinement; techniques such as targeted data augmentation or class-balancing loss functions could further improve sensitivity for the more advanced disease stages.

4.4 Analysis of the Contestable AI System

The model’s fallibility, particularly in intermediate and advanced disease stages, highlights the need for a human-in-the-loop system to identify and correct errors. We first evaluated our XMED safeguard, which is based on the hypothesis that

Table 1: Optimized hyperparameter configurations for the 1D-CNN model.

Hyperparameter	Value
Number of Convolutional Layers	5
Number of Fully Connected Layers	3
Activation Function	ReLU
Dropout Rate	0.5
Learning Rate	0.0003

Table 2: Classification performance of the 1D-CNN model on the test set.

Class	Precision	Recall	F1-Score	Support
Healthy	0.92	0.90	0.91	278
Stage 2	0.85	0.89	0.87	346
Stage 2.5	0.90	0.88	0.89	228
Stage 3	0.92	0.84	0.88	83
Accuracy			0.89	935
Macro Avg	0.90	0.88	0.89	935
Weighted Avg	0.89	0.89	0.89	935

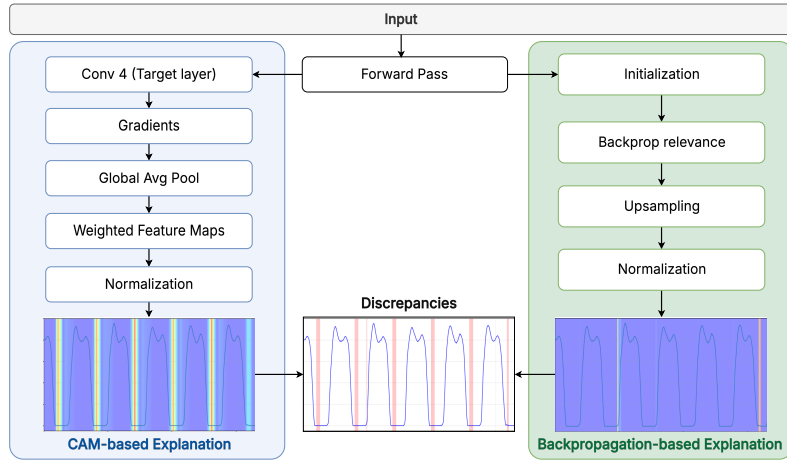


Fig. 4: Workflow overview of the XMED. The process compares CAM-based (Grad-CAM) and backpropagation-based (LRP) explanations to quantify model uncertainty. The input undergoes a forward pass to extract activations from the target convolutional layer. Grad-CAM computes weighted feature maps via gradient-based pooling, while LRP propagates relevance scores backward through the network. Both maps are normalized and compared to identify regions of high discrepancy, indicating divergent model explanations.

discrepancies between explanation methods can serve as a proxy for model uncertainty. To test this, we quantified the “high-discrepancy percentage” for a set of 30 test cases. The results confirm our hypothesis: misclassified cases exhibited a five-fold higher average discrepancy rate (7.56%) compared to correct predictions (1.45%). This demonstrates that attributional inconsistency is a reliable signal of model unreliability.

To calculate this metric, we generated two normalized explanation maps for each test sample: one using Grad-CAM [45] and another using LRP [20], as illustrated in Figure 4. We calculated the absolute difference at each timestep, flagging points where it exceeded a threshold of 0.5. These points were then merged into coherent high-discrepancy regions, and the final metric was the fraction of total timesteps falling within these regions.

With the safeguard validated, we investigated whether LLMs could use these signals for adjudication. We tested Llama 4 Scout (17B) [23] and GPT-4o (200B) [15] with identical prompts (Template 1) and settings. As summarized in Table 3, their performance profiles differed significantly. GPT-4o adopted a more conservative and reliable approach, correctly retaining all 24 correct predictions it reviewed. In contrast, Llama 4 was more interventionist; it successfully overturned two of the six incorrect predictions but also incorrectly overturned one correct case, suggesting different underlying reasoning processes that warrant further case-level analysis.

TEMPLATE 1: PROMPT TEMPLATE FOR CONTESTABLE GAIT ANALYSIS

SYSTEM MESSAGE: You are a helpful clinical decision support AI for Parkinson’s disease diagnosis using gait analysis. Always:

1. Think step-by-step before responding.
2. Justify your initial assessment and interpretation of gait metrics, referencing clinical guidelines or evidence when possible.
3. When finalization request is queried, you must finalize the decision (only answer: “Healthy”, “Stage 2”, “Stage 2.5”, or “Stage 3”) but you may overturn your prior assessment if, after reviewing all evidence, you are confident a different answer is correct.
4. Using your clinical analysis and justification, identify the potential reasons for any change in the final decision (e.g., specific gait abnormalities, asymmetries, variability, etc.) or in case of no change, justify why the initial assessment was correct. Then explain how these factors contribute to the final severity.
5. Provide accurate, current information using clinical gait analysis guidelines.
6. Cross-validate findings with multiple sources.
7. Reference sources for non-standard conclusions.
8. Maintain clarity with concise and straightforward responses.

USER MESSAGE:

1. **Prediction:** {class} (Confidence: {confidence}).
2. **Gait metrics:** {mean_stride_time}, {swing_percentage}, {stance_percentage}
3. **XAI analysis output:**
 - Discrepancy percentage: {discrepancy_percentage}%.
 - Continuous high-discrepancy regions: {[region_1], [region_2], ..., [region_n]}.

**Note that the numerical input values (in green) have been pre-calculated by the baseline model and XMED.*

Correct Prediction Case In a representative case of a correct classification (Figure 5a), the 1D-CNN identified a patient as Stage 0 (Healthy) with high confidence (0.821) and a correspondingly low XMED discrepancy of 0.8%. Both LLMs correctly upheld this initial prediction. As detailed in Figure 6, their justifications were grounded in clinical gait data, with both models observing that gait metrics were “within normal ranges.” Llama 4 further specified the “absence of Parkinsonian gait markers,” while GPT-4o correctly interpreted the minor discrepancy regions as not clinically significant, demonstrating a nuanced understanding of the XMED signal.

Incorrect Prediction Case A more revealing case involved a low-confidence (0.462) misclassification by the 1D-CNN, which was correctly flagged by a high XMED score of 6.5% (Figure 5b). In this instance, the LLMs diverged, as shown in Figure 7. Llama 4 successfully overturned the prediction, downgrading it from Stage 2.5 to Stage 2. It correctly reasoned that while the prolonged stance phase (65.1%) indicated impairment, the abnormality was insufficient to justify the higher severity rating. In contrast, GPT-4o retained the incorrect Stage 2.5 label, focusing on the clinical plausibility of the observed gait changes as compensatory mechanisms. It failed to differentiate the degree of this deviation, highlighting a more risk-averse adjudicative style that avoids overturning a prediction without overwhelming contradictory evidence.

Table 3: Number of classification cases by contestable LLMs with XMED support. The arrows ($\uparrow/\downarrow$) indicate higher/lower is better.

Model	Retain	Retain	Overturn	Overturn
	Correct $\uparrow$	Incorrect $\downarrow$	Correct $\downarrow$	Incorrect $\uparrow$
llama-4-scout-instruct (17B)	23	4	1	2
gpt-4o (200B)	24	5	0	1

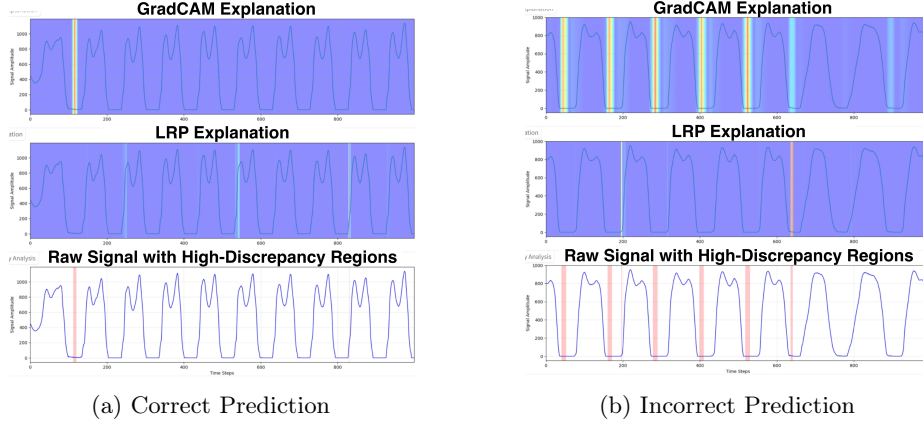


Fig. 5: XMED Visualization for (a) correct and (b) incorrect prediction cases.

In terms of computational efficiency, we observed a clear trade-off. Llama 4 was consistently faster, with a response time (RT) between 7 and 9 seconds, while GPT-4o was slower, taking between 11 and 15 seconds. However, GPT-4o produced more concise and direct justifications, with an output token (OT) count of approximately 499-562, compared to Llama 4’s more verbose outputs of 709-818 tokens. This suggests that while the smaller model offers lower latency, the larger model provides superior adjudication quality and more clinically-grounded reasoning.

4.5 Human-centered Evaluation Results

To assess the framework’s human-centered design in line with CAI principles, we evaluated Motion2Meaning on four key metrics. For a consistent and robust evaluation, these metrics were computed and averaged across the 30 test cases from the previous experiment, with the results summarized in Table 4.

Flesch Reading Ease (FRE) Llama 4 produced substantially more accessible justifications, achieving an FRE score of 51.6, which falls within the typical range for clinical documentation (50-70 [51,3]). In contrast, GPT-4o’s score of 34.73 places its output in the “difficult” range, reflecting a clear prioritization

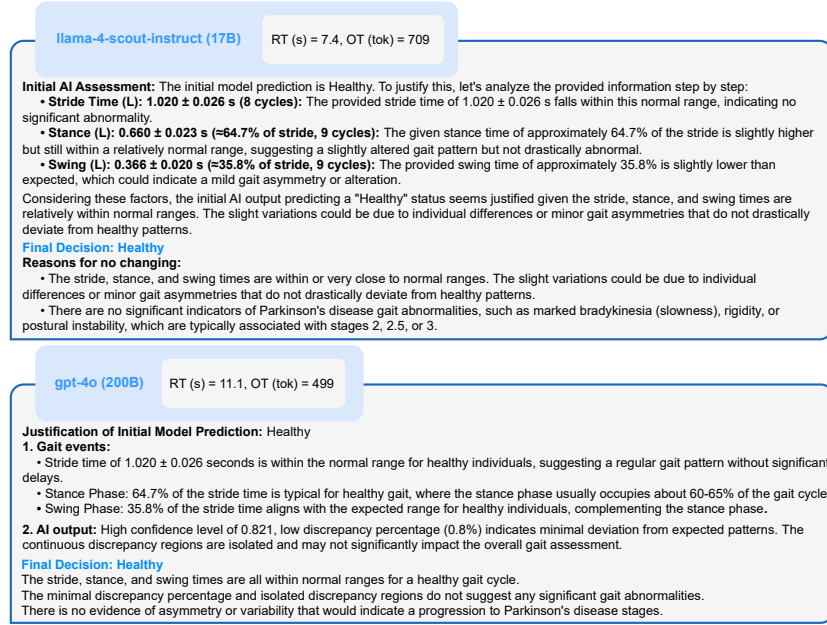


Fig. 6: Example responses from contestable LLMs in a Correct Prediction case. The final decision (in blue) indicates the model retains the correct initial prediction of baseline 1D-CNN. Output metrics: response time - RT (s) and number of output token - OT (tok).

of technical precision over linguistic simplicity. This highlights that readability is not an inherent model limitation but a highly tunable parameter. A simple directive within the prompt to target a specific readability level could readily align a model's output with the practical demands of a clinical workflow, making this a key consideration for system design.

Flesch-Kincaid Grade Level (FKGL) Both models produced outputs with an FKGL appropriate for clinical documentation, which typically targets a Grade 8-12 standard [51,3]. Llama 4 scored an 8.5, equivalent to a mid-8th-grade reading level, while GPT-4o scored a 10.52, corresponding to an early high school complexity. These results confirm that both models can generate explanations that are linguistically accessible to their intended clinical audience, with Llama 4 producing slightly more straightforward text. This finding is crucial, as it demonstrates that the complexity of the underlying AI reasoning does not have to result in an equally complex explanation for the end-user.

Clinical Grounding (CG) GPT-4o demonstrated superior factual grounding with a CG of 0.75, slightly outperforming Llama 4 at 0.72. Both models reliably

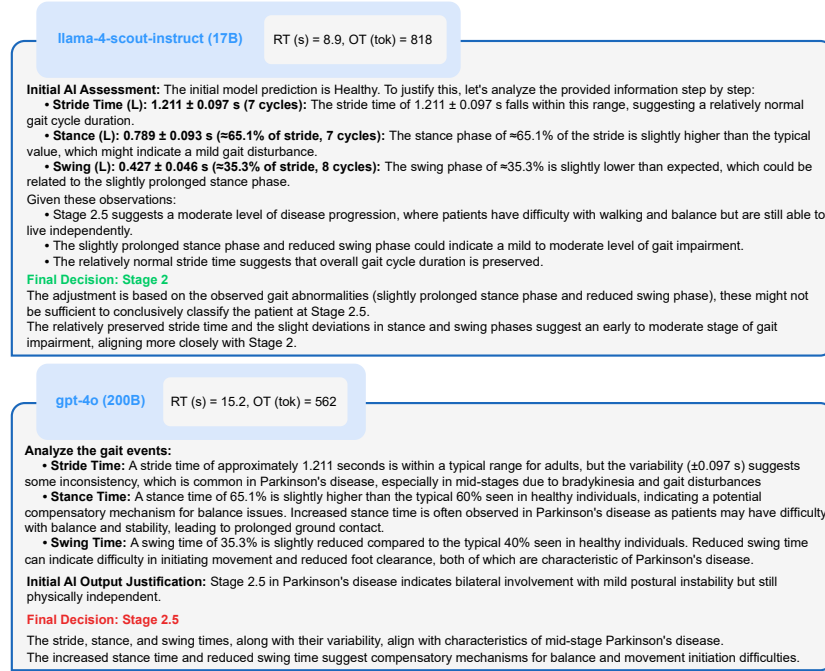


Fig. 7: Example responses from contestable LLMs in an Incorrect Prediction case. The final decision (in red) indicates the model retains the incorrect initial prediction of baseline 1D-CNN, while the final decision (in green) indicates the model overturns it. Output metrics: response time - RT (s) and number of output token - OT (tok).

reproduced most quantitative details from the input data, a critical capability in clinical domains where numerical precision is paramount. However, neither model achieved perfect accuracy. The persistence of minor hallucinations, even in state-of-the-art models, is a crucial finding that directly validates the necessity of a contestable framework. It underscores that human expertise is not merely beneficial but indispensable for verifying AI-generated claims before they can inform clinical decisions.

Self-Correction Accuracy (SCA) Llama 4 demonstrated superior responsiveness to contestation, achieving a higher SCA score of 0.33 compared to GPT-4o's 0.17. This suggests the smaller model is more adaptive and willing to revise its initial assessment in light of contradictory evidence. In contrast, GPT-4o's lower score reflects a more conservative, risk-averse behavior, where it tends to default to the baseline model's prediction. This finding reveals a crucial trade-off: while larger models may offer greater factual grounding, smaller models might be

Table 4: Human-centric evaluation metrics. The arrows ($\uparrow/\downarrow$) indicate whether higher or lower values are better.

Model	FRE $\uparrow$	FKGL $\downarrow$	CG $\uparrow$	SCA $\uparrow$
llama-4-scout-instruct (17B)	51.60	8.50	0.72	0.33
gpt-4o (200B)	34.73	10.52	0.75	0.17

more amenable to the corrective feedback that is central to a truly collaborative human-AI system.

Our results indicate that Llama 4 produced more readable and adaptive explanations, whereas GPT-4o demonstrated superior factual grounding at the cost of higher linguistic complexity. This reveals a critical design trade-off between factual veracity and adaptive reasoning. The selection of an LLM is therefore not a simple technical optimization but a decision that fundamentally shapes the nature of the human-AI partnership, balancing the need for a reliable adjudicator against that of a collaborative and correctable partner.

5 Discussion

Our work demonstrates the feasibility of Motion2Meaning, a framework that successfully integrates AI-powered gait analysis with a contestable, human-in-the-loop interface. The findings confirm that it is possible to build systems that provide objective insights into motor symptoms while preserving essential clinical oversight through a structured and auditable workflow. We structure our discussion around two key themes. First, we reflect on the potential of AI-powered wearable gait analysis to transform PD care, considering our predictive model’s performance and limitations. Second, we analyze the challenges and future directions for developing truly human-centered CAI, drawing specific insights from the performance of our XMED safeguard and LLM-based interaction components.

5.1 The Potential of AI-Powered Wearable Gait Analysis in PD Care

Our results, showing that a 1D-CNN can effectively classify disease severity from raw gait signals, reinforce the potential of wearable sensors to shift patient assessment from episodic clinical snapshots to continuous, longitudinal monitoring. This objective data stream offers a powerful complement to subjective tools like the UPDRS, enabling clinicians to more precisely track therapy response, detect subtle motor fluctuations, and quantify changes in fall risk. The ultimate promise of this technology is a move toward more personalized and proactive treatment adjustments that can tangibly improve patients’ quality of life.

However, a purely gait-focused approach has inherent limitations. The nuances of intermediate disease stages and the multi-system nature of PD, which includes significant non-motor symptoms like cognitive impairment and sleep disturbances, underscore that gait is only one piece of a complex clinical picture. Therefore,

the true value of this technology lies in augmenting, not replacing, holistic clinical judgment. The critical next step is to move beyond a unimodal biomarker toward a comprehensive digital phenotype of PD. Future work should focus on multi-modal fusion, integrating gait data with other streams like speech analysis and sleep tracking. This approach is essential not only for improving predictive accuracy but for capturing the true syndromic nature of the disease, leading to more robust and clinically relevant AI models.

5.2 Toward Human-Centered Contestable AI in Healthcare

The fallibility of our predictive model, despite its reasonable performance, underscores the critical need for robust human-in-the-loop systems. Our work addresses this by creating a multi-layered verification framework where the XMED module acts as an automated safeguard and the LLM serves as an interactive adjudicator. This design directly confronts the fundamental challenge of balancing automation’s efficiency with the necessity of human oversight. However, our findings also illuminate several key challenges that must be addressed to advance the development of truly effective and collaborative CAI.

Technical and Methodological Limitations Our evaluation confirms that explanation discrepancies can effectively signal prediction errors, validating XMED’s potential as an automated safeguard. The scope of this validation, however, has two key considerations. First, the XMED flagging mechanism relied on an empirically derived discrepancy threshold. While effective for this study, its optimal calibration for broader clinical use is a natural next step for future work. Second, our framework was developed and validated exclusively within the context of PD, which served as a motivating use case. Its applicability and potential modifications for other diagnostic domains remain an important area for further exploration.

Furthermore, our experiments with LLMs highlight an interesting performance trade-off between the more adaptive reasoning of Llama 4 and the more factually grounded but conservative style of GPT-4o. This suggests that the selection of a model for adjudication involves balancing different priorities. A crucial open question, therefore, is whether this trade-off is inherent to LLM-based adjudication or is an artifact of using general-purpose models. The next logical step is to systematically evaluate models specifically fine-tuned on medical corpora, such as the Med-PaLM [46,47] and BioMistral [19] families, and to explore the capabilities of next-generation generalist models like GPT-5 [50], to determine if they can resolve the tension between adaptability and reliability. The modular design of our framework is a key strength in this regard, as it readily allows for the substitution of the predictive model, the LLM adjudicator, or the underlying dataset. This flexibility ensures that the core principles of our contestable system can be adapted and refined as new models and new clinical use cases emerge.

Challenges in Clinical Validation A primary challenge arising from our study is the need for a comprehensive clinical evaluation. While our technical validation

provides encouraging proof-of-concept results, the true measure of our framework’s success lies in its real-world utility. This necessitates a formal clinician-in-the-loop pilot study to move beyond automated metrics and assess the system’s impact on actual clinical practice. Such a study would involve close collaboration with neurologists to evaluate the correctness, clarity, and actionability of the LLM’s outputs and to measure key workflow metrics like review time and contest rates. Gathering this contextual, human-centered evidence through detailed case studies is the essential next step for responsibly translating this research prototype into a validated clinical tool.

Future Direction on Evaluation Metrics A truly human-centered approach requires a paradigm shift in evaluation. This involves conducting formal pilot studies with expert clinicians to move beyond accuracy to a suite of utility metrics. These can be divided into automated assessments of the AI’s output and observational measures of user interaction. For instance, an automated metric like Clinical Terminology Grounding (CTG) could assess if an explanation is grounded in professional language by calculating the percentage of sentences containing terms from a predefined clinical lexicon. A high score would indicate reasoning more plausible to an expert. Observational metrics could quantify interaction efficiency through measures like Time to Decision (TTD), the duration until a user validates or contests a finding, and Interaction Length (IL), the number of conversational turns needed to reach that decision. Beyond these, our evaluation could be extended by adapting established metrics of user reliance, such as the Relative AI Reliance (RAIR), Relative Self-Reliance (RSR), and Appropriateness of Reliance (AoR) proposed in [44], as well as the broader range of measures collected in [22,14].

6 Conclusion

This work demonstrates that Motion2Meaning unites SOTA AI gait analysis with a contestable, human-centered framework to deliver objective, accountable, and clinically viable interpretations of Parkinsonian motor symptoms. By combining accurate 1D-CNN classification with the XMED safeguard for uncertainty detection and the LLM-driven “contest and justify” workflow, the system ensures that clinical expertise remains central to decision making. Beyond advancing PD care, our results highlight a broader design principle: high-stakes medical AI must not only explain but also enable contestation, creating systems that are transparent, auditable, and aligned with regulatory requirements. Looking forward, integrating multimodal data, expanding to diverse populations, and tailoring domain-specific LLMs will further strengthen this approach. As global frameworks increasingly mandate explainability and contestability, Motion2Meaning provides a concrete step toward trustworthy AI that augments rather than replaces human judgment, setting a foundation for safer and more responsible deployment of AI in healthcare.

References

1. Alharthi, A.S.: Interpretable machine learning comprehensive human gait deterioration analysis. *Frontiers in Neuroinformatics* **Volume 18 - 2024** (2024)
2. AlMahadin, G., Lotfi, A., Zysk, E., Siena, F.L., Carthy, M.M., Breedon, P.: Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. *BMC Neurology* **20**(1), 419 (Nov 2020)
3. Challener, D.W., Wen, A., Fan, J.W., Liu, H., O'Horo, J., Nyman, M.: Flesch-Kincaid Grade Level Readability Scores to Evaluate Readability of Clinical Documentation During an Electronic Health Record Transition. *Advances in Health Information Science and Practice* **1**(1), VBWY7913 (2025)
4. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018)
5. Daniore, P., Nittas, V., Haag, C., Bernard, J., Gonzenbach, R., von Wyl, V.: From wearable sensor data to digital biomarker development: ten lessons learned and a framework proposal. *npj Digital Medicine* **7**(1), 161 (Jun 2024)
6. de Deus Fonticoba, T., Santos García, D., Macías Arribí, M.: Inter-rater variability in motor function assessment in parkinson's disease between experts in movement disorders and nurses specialising in pd management. *Neurología (English Edition)* **34**(8), 520–526 (2019)
7. El Maachi, I., Bilodeau, G.A., Bouachir, W.: Deep 1d-convnet for accurate parkinson disease detection and severity prediction from gait. *Expert Systems with Applications* **143**, 113075 (Apr 2020)
8. European Union: Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (2024), accessed: 2025-02-07
9. Flesch, R.: How to write plain english. University of Canterbury (1979)
10. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* **101**(23), e215–e220 (2000)
11. Herrera, F.: Making Sense of the Unsensible: Reflection, Survey, and Challenges for XAI in Large Language Models Toward Human-Centered AI. *arXiv preprint arXiv:2505.20305* (2025)
12. Hoehn, M.M., Yahr, M.D.: Parkinsonism: onset, progression, and mortality. *Neurology* **17**(5), 427–427 (1967)
13. Hoffman, S.L., Schmiedmayer, P., Gala, A.S., Wilkins, K.B., Parisi, L., Karjagi, S., Negi, A.S., Revlock, S., Coriz, C., Revlock, J., et al.: Comprehensive real time remote monitoring for parkinson's disease using quantitative digitography. *npj Parkinson's Disease* **10**(1), 137 (2024)
14. Hsiao, J.H.w., Ngai, H.H.T., Qiu, L., Yang, Y., Cao, C.C.: Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI). *arXiv preprint arXiv:2108.01737* (2021)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o System Card. *arXiv preprint arXiv:2410.21276* (2024)

16. Jankovic, J.: Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry* **79**(4), 368–376 (2008)
17. Janssen Daalen, J., van der Heiden, M., Meinders, M., Post, B.: Motor symptom variability in parkinson's disease: Implications for personalized trial outcomes? *Movement Disorders* **40**(5), 975–979 (2025)
18. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep. (1975)
19. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains (2024)
20. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: The LRP Toolbox for Artificial Neural Networks. *Journal of Machine Learning Research* **17**(114), 1–5 (2016)
21. Li, J., Yang, Y., Zhang, R., Lee, Y.c.: Overconfident and Unconfident AI Hinder Human-AI Collaboration. arXiv preprint arXiv:2402.07632 (2024)
22. Lopes, P., Silva, E., Braga, C., Oliveira, T., Rosado, L.: XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences* **12**(19) (2022)
23. Meta: The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Meta AI Blog (2025)
24. Mobbs, R.J., Perring, J., Raj, S.M., Maharaj, M., Yoong, N.K.M., Sy, L.W., Fonseka, R.D., Natarajan, P., Choy, W.J.: Gait metrics analysis utilizing single-point inertial measurement units: a systematic review. *mHealth* **8**(0) (2021)
25. Monje, M.H., Grosjean, S., Srp, M., Antunes, L., Bouça-Machado, R., Cacho, R., Domínguez, S., Inocentes, J., Lynch, T., Tsakanika, A., et al.: Co-designing digital technologies for improving clinical care in people with parkinson's disease: What did we learn? *Sensors* **23**(10), 4957 (2023)
26. Motahari-Nezhad, H., Fgaier, M., Mahdi Abid, M., Péntek, M., Gulácsi, L., Zrubka, Z.: Digital biomarker-based studies: Scoping review of systematic reviews. *JMIR Mhealth Uhealth* **10**(10), e35722 (Oct 2022)
27. Movement Disorder Society Task Force On Rating Scales For Parkinson's Disease: The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorders* **18**(7), 738–750 (2003)
28. Naimi, S., Bouachir, W., Bilodeau, G.A.: 1d-convolutional transformer for parkinson disease diagnosis from gait (2023)
29. Navita, Mittal, P., Sharma, Y.K., Rai, A.K., Simaiya, S., Lilhore, U.K., Kumar, V.: Gait-based parkinson's disease diagnosis and severity classification using force sensors and machine learning. *Scientific Reports* **15**(1), 328 (Jan 2025)
30. Nguyen, H., Clement, T., Nguyen, L., Kemmerzell, N., Truong, B., Nguyen, K., Abdelaal, M., Cao, H.: Langxai: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. In: Larson, K. (ed.) *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. pp. 8754–8758. International Joint Conferences on Artificial Intelligence Organization (8 2024), demo Track
31. Nguyen, H., Rahimi, A., Whitford, V., Fournier, H., Kondratova, I., Richard, R., Cao, H.: Heart2Mind: Human-Centered Contestable Psychiatric Disorder Diagnosis System using Wearable ECG Monitors. arXiv preprint arXiv:2505.11612 (2025)
32. Nguyen, H.T.T., Nguyen, L.P.T., Cao, H.: Xedgeai: A human-centered industrial inspection framework with data-centric explainable edge ai approach. *Information Fusion* **116**, 102782 (2025)

33. Nguyen, L.P.T., Nguyen, H.T.T., Cao, H.: Odexai: A comprehensive object detection explainable ai evaluation. In: Braun, T., Paaßen, B., Stolzenburg, F. (eds.) KI 2025: Advances in Artificial Intelligence. pp. 118–133. Springer Nature Switzerland, Cham (2026)
34. Nguyen, P.T.L., Do, T.H.: ConGaIT: A Clinician-Centered Dashboard for Contestable AI in Parkinson’s Disease Care. *Mensch und Computer 2025 - Workshopband* (2025)
35. Park, H., Shin, S., Youm, C., Cheon, S.M.: Deep learning-based detection of affected body parts in parkinson’s disease and freezing of gait using time-series imaging. *Scientific Reports* **14**(1), 23732 (Oct 2024)
36. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models (2018)
37. Prasanth, H., Caban, M., Keller, U., Courtine, G., Ijspeert, A., Vallery, H., von Zitzewitz, J.: Wearable sensor-based real-time gait detection: A systematic review. *Sensors* **21**(8) (2021)
38. Prisco, G., Pirozzi, M.A., Santone, A., Esposito, F., Cesarelli, M., Amato, F., Donisi, L.: Validity of wearable inertial sensors for gait analysis: a systematic review. *Diagnostics* **15**(1), 36 (2024)
39. Rashnu, A., Salimi-Badr, A.: Integrative deep learning framework for parkinson’s disease early detection using gait cycle data measured by wearable sensors: A cnn-gru-gnn approach (2024)
40. Renggli, D., Graf, C., Tachatos, N., Singh, N., Meboldt, M., Taylor, W.R., Stieglitz, L., Schmid Daners, M.: Wearable inertial measurement units for assessing gait in real-world environments. *Frontiers in physiology* **11**, 90 (2020)
41. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 1135–1144. KDD ’16, Association for Computing Machinery, New York, NY, USA (2016)
42. Romeo, G., Conti, D.: Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. *AI & SOCIETY* (Jul 2025)
43. Salchow-Hömmen, C., Skrobot, M., Jochner, M.C.E., Schauer, T., Kühn, A.A., Wenger, N.: Review—emerging portable technologies for gait analysis in neurological disorders. *Frontiers in Human Neuroscience* **Volume 16 - 2022** (2022)
44. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G.: Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. p. 410–422. IUI ’23, Association for Computing Machinery, New York, NY, USA (2023)
45. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
46. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (Aug 2023)
47. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward Expert-Level Medical Question Answering with Large Language Models. *Nature Medicine* **31**(3), 943–950 (2025)

48. The European Parliament and the Council of the European Union: Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, L 119/1 (2016), cited for Article 22.
49. Truong, V.B., Nguyen, H., Nguyen, V.T.K., Nguyen, Q.K., Cao, Q.H.: Towards Better Explanations for Object Detection. In: Yamkoğlu, B., Buntine, W. (eds.) Proceedings of the 15th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 222, pp. 1385–1400. PMLR (Nov 2024)
50. Wang, S., Hu, M., Li, Q., Safari, M., Yang, X.: Capabilities of GPT-5 on Multimodal Medical Reasoning. arXiv preprint arXiv:2508.08224 (2025)
51. Wu, D.T., Hanauer, D.A., Mei, Q., Clark, P.M., An, L.C., Lei, J., Proulx, J., Zeng-Treitler, Q., Zheng, K.: Applying Multiple Methods to Assess the Readability of a Large Corpus of Medical Documents. *Studies in health technology and informatics* **192**, 647 (2013)
52. Yin, W., Zhu, W., Gao, H., Niu, X., Shen, C., Fan, X., Wang, C.: Gait analysis in the early stage of parkinson’s disease with a machine learning approach. *Frontiers in Neurology* **Volume 15 - 2024** (2024)
53. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 295–305. FAT* ’20, Association for Computing Machinery, New York, NY, USA (2020)