SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

# Multimodal Attention Transformer Framework for Parkinson's Disease Detection Using Gait, Speech, Wearable Sensors, and Auxiliary Neuroimaging

D. Geetha[1] , M. Sunil Kumar[2]*

[1]Research Scholar, Department of CSE School of Computing Mohan Babu University, Tirupati, AP, India geethakrish18@gmail.com
[2]Professor & Dean-P&M, Department of CSE School of Computing Mohan Babu University, Tirupati, AP, India sunilmalchi1@gmail.com

**Abstract**
Early detection of Parkinson's disease (PD) is complicated by the insidious, heterogeneous nature of early symptoms that are incompletely measured with a single biomarker. This study introduces a multimodal artificial intelligence framework that combines gait, speech and wearable sensor data as primary digital biomarkers supported by an auxiliary validation branch for Parkinson's Progression Markers Initiative (PPMI) neuroimaging. These models included modality-specific and fusion-based models that were developed using publicly available datasets. Locomotor features were extracted from gait signals, acoustic biomarkers from speech data, movement variability metrics from wearable signals and imaging-derived representations from neuroimaging data. Several temporal regression models were explored: CNN, LSTM, CNN-LSTM, Transformer,CNN-Transformer and a multimodal attention-based transformer model. The gait-speech-wearable multimodal Transformer proposed here provided 93.1% accuracy and 0.965 AUC, which was superior to single-modality and conventional deep learning models with body-worn sensors alone. The highest performances were achieved by using the imaging-supported multimodal Transformer, with 94.3% accuracy and 0.973 AUC. Explainability of attention demonstrated that complementary diagnostic features from gait, speech and imaging were provided by wearable data. This approach may inform future decision-support systems in screening and monitoring for Parkinson's disease, but clinical deployment would require validation against matched prospective multimodal datasets.

**Keywords:** Parkinson's disease detection; multimodal artificial intelligence; medical imaging-supported AI; digital biomarkers; gait analysis; speech biomarkers; wearable sensors; Transformer model; attention-based fusion; explainable AI.

## 1. Introduction

Parkinsons disease (PD) is among the most rapidly growing global neurological disorders as well as a leading clinical, social and economic burden. Collective global estimates show that 2019 there were more than 8.5 million people living with PD worldwide, and updated modelling studies suggest this would rise to nearly 25.2 million by 2050 as a result of population ageing, growth and changing patterns of disease prevalence [1] [2]. PD clinically is characterized primarily as motor symptoms of bradykinesia, resting tremor, rigidity and postural instability/gait impairment along with constellation of non-motor symptoms such sleep disturbance, autonomic dysfunction, cognitive decline, mood-related symptomatology and sensory disturbances [3]. These symptoms are heterogeneous in appearance across patients and disease stages and consequently difficult to be recognized early in the routine clinical practice. As the burden of PD grows, so does the demand on healthcare systems, caregivers and families; managing this degenerative disease is a long-term task requiring adjustment of medication, rehabilitation of motor complications and support with activities of daily living [1]. Thus, it is important for timely clinical management, risk stratification and monitoring of patients as well as enrollment in clinical trials testing disease-modifying interventions that early and accurate detection [4].

Despite significant advances in delineating the clinical and biological features of PD, diagnosis still largely relies on expert neurological examination. Standard evaluation relies on history and clinical examination focusing primarily on overt motor signs, supplemented by rating scales that reflect more global activity than subtle features like bradykinesia or fluctuations outside the clinic [3] [5]. During the prodromal and early motor phases, when symptoms such as hyposmia, constipation, sleep disturbance, and mild movement changes would be difficult to differentiate from non-specific ageing or other neurological conditions [3], diagnostic uncertainty is particularly high. Distinguishing idiopathic PD from atypical parkinsonian syndromes, such as multiple system atrophy, progressive supranuclear palsy and corticobasal degeneration can also be challenging in the first few years of disease progression [5]. While biological staging frameworks (BSF) and biomarker-based definitions of early disease are on the rise, we continue to lack scalable and routinely deployable tools which allow for early and objective assessment [6]. This gap in diagnostic capabilities has prompted efforts to develop digital biomarkers that can aid clinical decisions based on continuous, quantitative, and objective real-world measurements.

Short time-frame clinic-based examinations may not capture relevant disease-related changes, and digital health technologies provide a unique opportunity to do so. Wearable sensors, smartphone-based assessments, and inertial measurement units can obtain movement and activity data in real environments of daily living, greatly addressing gait [7], [8]. The study of gait has been particularly useful because PD-associated alterations in stride length, step timing, walking speed, and gait variability have remained consistent indicators of motor impairment and disease progress [8]. Similar to previous research [9], this study found that speech analysis not only provides a non-invasive digital marker, as PD may impact vocal loudness and pitch stability, articulation, phonation and speech rhythm. Work on PD classification from gait [8], speech [9] and wearable sensor signals [10] has mostly been driven using machine learning and deep learning methods. Nonetheless, much of the current AI-based work continues to rely on a unimodal basis and thus may fail when evaluating models across different populations, recording conditions, device types, medication states and disease stages [10]. Such single-modality models suffer from lack of generalizability and clinical readiness.

Since gait, speech and wearable sensor data reflect complimentary aspects of PD related impairment it makes sense to utilisha multimodal approach. Most of the gait signals can be seen as indicative of locomotion dysfunction, while vocal and articulatory characteristics can be inferred from speech signals, and wearable sensors provides continuous capture of motor behavior as they happen in naturalistic environments. Thus, a combination of these modalities could yield a more comprehensive disease representation than any one biomarker on its own. Integrated multimodal machine learning approaches based on clinical, imaging, genetic and sensor-related data provide better performance in predicting PD than isolated modalities [11]. Findings from a separate multimodal deep learning and explainable AI study, also using the Parkinson's Progression Markers Initiative data set, further suggest that joint multi-modal classification may yield us both clinically meaningful interpretability as well as boost early PD detection sensitivity [12]. Moreover, contemporary reviews have pointed out the increasing importance of multimodal data fusion, Transformer-based models and heterogeneous digital biomarkers in PD detection and review [13]. However, current multimodal studies tend to emphasize only a few combinations of clinical, imaging, genetic, or task-specific variables compared with the number of deep learning frameworks that have been fully integrated for gait kinematics; acoustic speech features; and continuous wearable sensor streams.

This is an important aspect for multimodal PD detection as those structures can model long-range temporal dependencies in sequential data. This is especially pertinent in long rhythm patterns like gait cycles, speech signals and wearable sensor streams − diseases manifest more so across longer temporal patterns rather than isolated segments of those signals. Recent work has demonstrated the application of Transformer-based and ConvNet-Transformer models for PD-related gait as well as freezing-of-gait (FOG) detection, which denotes a one-dimensional medical and wearable sensor signal modelling problem [14], [15]. In addition, they also find an adjacent area of research in wider biomedical AI literature that supports the idea that multimodal learning frameworks can incorporate heterogeneous medical data sources resulting in improved clinical prediction tasks [16]. Yet, the simultaneous processing of gait, speech and multi-sensory data for PD detection via attention-guided Transformer fusion has been little studied. This can potentially define a methodological void to develop an interpretable multimodal AI framework that can learn modality-specific and cross-modal disease representations.

In addition, neuroimaging also plays a vital role in the research as well as for differential diagnosis of Parkinson's disease. Structural and quantitative information about brain regions affected by neurodegeneration are provided for MRI, while dopamine transporter (DaT) imaging (DaTscan & SPECT), support the presence of presynaptic dopaminergic deficit in the Parkinsonian disorders [17], [18]. Neuroimaging-derived features have supported machine learning studies using imaging and clinical data

for automated Parkinson disease diagnosis and early detection [19]. Also, image based models for the classification of Parkinson's disease have recently been explored in [20] as well among other recent deep learning studies on MRI. Resources available through public access, such as the Parkinsons Progression Markers Initiative which exposes longitudinal clinical, imaging, genetic, biospecimen and biomarker data are appropriate for multi-modal AI frameworks validated using imaging alone [17]. Neuroimaging is not a surrogate for gait, speech and wearable biosensors biomarkers in this work. A separate auxiliary branch validates that the imaging-derived representations can complement and reinforce the main framework of inferring Parkinson's disease based on digital biomarkers.

This study fills the aforementioned gap by proposing a multimodal artificial intelligence framework to fuse gait, speech and wearable sensor data for early and accurate detection of Parkinsons across diverse patient populations. The main framework fuses three distinct digital measurement streams: quantitative gait features, acoustic speech features, wearable accelerometer and gyroscope signals. In order to reach better relevance for medical imaging-supported validation, Parkinsons Progression Markers Initiative (PPMI) neuroimaging data would be the auxiliary secondary path of a summary analysis [17], [18] which may strengthen inferences through MRI and when available dopamine transporter imaging. The CNN, LSTM, CNN-LSTM, Transformer models are positioned for comparative comparison and the proposed advanced architecture is the multimodal attention-based Transformer framework (MC-ATF). This work facilitates the development of generalizable, interpretable artificial intelligence tools for detecting Parkinsons disease in earlier states by incorporating multimodal fusion, explainability and optional imaging-supported validation into the study design. The following section discusses related work on digital biomarkers, neuroimaging-based Parkinson's disease evaluation, multimodal artificial intelligence and Transformer models to highlight the methodological gaps filled by the proposed framework.

## 2. Related Work / Literature Review

Over the past years, Parkinson's disease research has shifted toward objective digital biomarkers owing to the fact that most motor and non-motor changes are challenging to objectively ascertain during a brief clinical evaluation. Gait is of particular interest as walking impairment is not only one of the most disabling features of Parkinson's disease, but also among the most strongly associated with mobility loss, increased fall risk, reduced independence and poorer quality of life. Previous clinical and biomechanical studies indicated that gait impairment observed in Parkinsons disease includes reduced gait speed, shorter step length, rhythmicity disturbance, turning deficits, freezing episodes as well as increased stride-to-stride variability [21], [22]. Based on these observations, the foundations for digital assessment of gait via wearable sensors and signal processing were created. Subsequent studies provided evidence that long-duration mobility monitoring can assess turning measures, gait variability, and movement characteristics in daily life that would not be detected in a short clinic visit [23]. Fortunately, recent reviews noted that while gait digital biomarkers make disease-risk estimation and progression monitoring possible as well as fall prediction and treatment-response assessments [24], the protocols must be standardized and rigorously externally validated across cohorts. This means gait alone offers an important but insufficient perspective on Parkinson's: locomotor changes do not capture speech impairment, tremor fluctuation, drug-related fluctuations and more global functional decline.

More recently, studies using smartphone and wearable sensors expanded this field by enabling repeated measurement in homes and free-living environments. The mPower study was one of the first large app-based Parkinson's disease studies to show that data on gait, voice, tapping and self-reported symptoms could be collected using smartphone-based participation [25]. Wearable sensor studies, such as the CIS-PD work, demonstrated that machine learning models can detect bradykinesia and tremor in movement data collected using body-worn sensors during clinically rated tasks [26]. Recent multicentre work from WATCH-PD demonstrated that a smartphone and smart watch measures capture manifestations of early Parkinson's disease; arm swing, tremor-related measures and finger-tapping performance all supported future digital endpoint design [27]. Platforms such as the Roche Parkinson's Disease Mobile Application for smartphones and smart watches have also demonstrated reliability and clinical validity for remote monitoring of early features of Parkinson's disease [28]. Passive sensor data has also been applied to estimate symptom severity such as medication state, dyskinesia and tremor from free-living datasets, and in crowd sourced challenges by making open reproducible code available [29], which highlight the challenge of creating robust models when faced with real world noise. Additionally, inertial sensing via smart watch has been proven useful to enable continuous monitoring of tremor and dyskinesia in the community providing further support for wearable motor biomarkers [30]. Taken together, these studies establish the clinical validity of wearable and smartphone-derived signals, yet also demonstrate that sensor-only strategies are still susceptible to device placement, adherence, data quality and population heterogeneity.

Parkinson's disease detection based on speech shares the same evolution timeline as gait and wearable monitoring. Voice impairments are prevalent in Parkinsons disease (PD) and can lead to low vocal loudness, audible breathiness, monopitch, imprecise articulation, abnormal phonation quality and decreased prosodic variation. Dysphonia measures and acoustic features of speech can separate PD from healthy control speech [31], [32] and support remote assessment of disease severity through passive acoustics. Later studies built on this work by provided comparisons between acoustic feature extraction methods, features based upon the tunable Q-factor wavelet transform, and traditional machine learning classifiers, for Parkinson's disease classification [33]. Recent deep learning research is not limited to hand-crafted acoustic descriptors, but interprets representations directly from speech signals with convolutional and end-to-end architectures [34]. There are systematic reviews of deep learning approaches applied to speech [35], suggesting that in the space of CNNs, recurrent models, and transfer learning have been most explored but with consistently limited external validation focusing on small datasets, language bias, variability across recording conditions and privacy. Machine learning methods applied on voice biomarkers still demonstrate that speech records can aid to classify Parkinson's disease, but the reported performance is highly dependent on feature dimensionality, recording protocol, and validation design [36]. As such, speech is a convenient and non-invasive one, however, pure speech models may not fully reflect the motor and functional level phenotype of Parkinson's disease.

The restrictions of single-modality models have prompted the use of multi-modal machine learning. Early predictions of Parkinson's disease have shown that the models are more interpretable, especially when using explainable machine learning methods [37], and multimodal time-series studies have expressed how heterogeneous data integration can be advantageous. Multimodal features have been used with deep learning architectures; integrating biological, imaging and clinical information outperformed the isolated data streams for classification [38]. Machine learning reviews within the Parkinson's Progression Markers Initiative have emphasized that the data set includes longitudinal and multimodal information of many types, but much of this structure is poorly utilized in current machine learning analyses [39]. This is a major point because Parkinson's disease is clinically heterogeneous and models trained on narrow feature sets may not generalize to broader representations of the disease. Both of these important considerations are further supported in wider literature, since multimodal machine learning is a natural method to use when clinical diagnosis often relies on a combination of multiple input data modalities such as imaging, time-series signals, tabular clinical data and patient history [40]. Likewise, multimodal biomedical artificial intelligence has been characterized as a potentially robust strategy to enhance prediction, diagnosis and monitoring by combining complementary information from different biomedical sources [16]. These findings underpin the main hypothesis of this work that gait data, speech data and wearable sensor data are better analyzed in a unified framework rather than separately.

They have also been utilized in diagnosis and biomarker development for Parkinson's disease through the use of medical imaging. SPECT imaging with dopamine transporter and DaTscan can give information on presynaptic dopaminergic deficit, which may help in the differentiation between Parkinsonian syndromes and non-degenerative movement disorders [18],[41]. Because MRI-based techniques offer the structural and quantitative features of regional neurodegeneration, recent investigation of imaging approaches has been introduced using machine learning and deep learning to automatically classify Parkinson's disease [19], [20]. This is particularly pertinent given that the Parkinson's Progression Markers Initiative (PPMI) integrates clinical, imaging, genetic, biospecimen, sensor and longitudinal biomarker data in an observational research framework [17]. On the contrary, image-only models either ignore daily motor behaviour variability and speech impairments or how functional fluctuations associate with neurocognitive metrics. Thus, a multimodal strategy, with neuroimaging-derived representations integrated into non-image digital biomarkers such as gait, speech and/or wearable sensor data may represent a more robust design.

Advances in deep learning provide additional support for the adoption of Transformer-based models when it comes to detection of Parkinson's disease. The original Transformer architecture incorporated self-attention, a mechanism for expressing relationships over long sequences without dependence on recursivity [42]. This is pertinent to Parkinsons because gait cycles, periods of speech-specific patterns, episodes of tremor, bradykinesia and dyskinesia may extend over long temporal windows. 1D CNNs are also effective for time-series analysis, but surveys of Transformer models indicate that attention-based architectures are useful when long-range dependency, variable temporal behavior and interaction between multivariate signals [43]. For Parkinson's disease this entails that a Transformer-based model has the capacity to actually learn both intra-modality temporal structure and cross-modality relationships among gait, speech and wearables streams. But currently, the majority of Parkinson's disease studies are still based on single-modality deep learning or limited multimodal combinations, and relatively few frameworks connect gait kinematics with acoustic speech features and continuous wearable sensor

signals through an attention-guided architecture Thus, it presents an obvious methodological void for a multimodal attention-based Transformer framework to systematically compare CNN, LSTM, CNN-LSTM and Transformer-based models (e.g., single-modality and multimodal variants) on a common evaluation design. The study proposed accordingly benefits from the advantages of previous digital biomarker studies while concurrently resolving the requirement for sound fusion, interpretability and generalisability across diverse patient cohorts. Motivated by these gaps, the current work presents a multimodal attention-based Transformer framework to help support early and accurate detection of Parkinson's disease through gait, speech and wearable sensor data using a unified deep learning pipeline.

## 3. Methodology

In the current study, a medical imaging assisted multimodal artificial intelligence framework for the detection of Parkinson's Disease is used while keeping the primary goal of incorporating gait, speech and wearable sensor data. The methodology is optimized for a medical imaging journal with an ancillary branch for neuroimaging validation of the method, incorporating data from the Parkinson's Progression Markers Initiative (PPI) (MRI and dopamine transporter (DT) imaging, where available). The main experimental setting is still based on the non-image digital markers (gait, speech and signals from wearable sensors); neuroimaging is employed to enhance clinical relevance, secondary validation, and imaging-derived contextual evidence. This alignment enables the study to continue its trajectory towards the original goal but also bring the framework to a multimodal biomedical artificial intelligence supported by medical imaging.

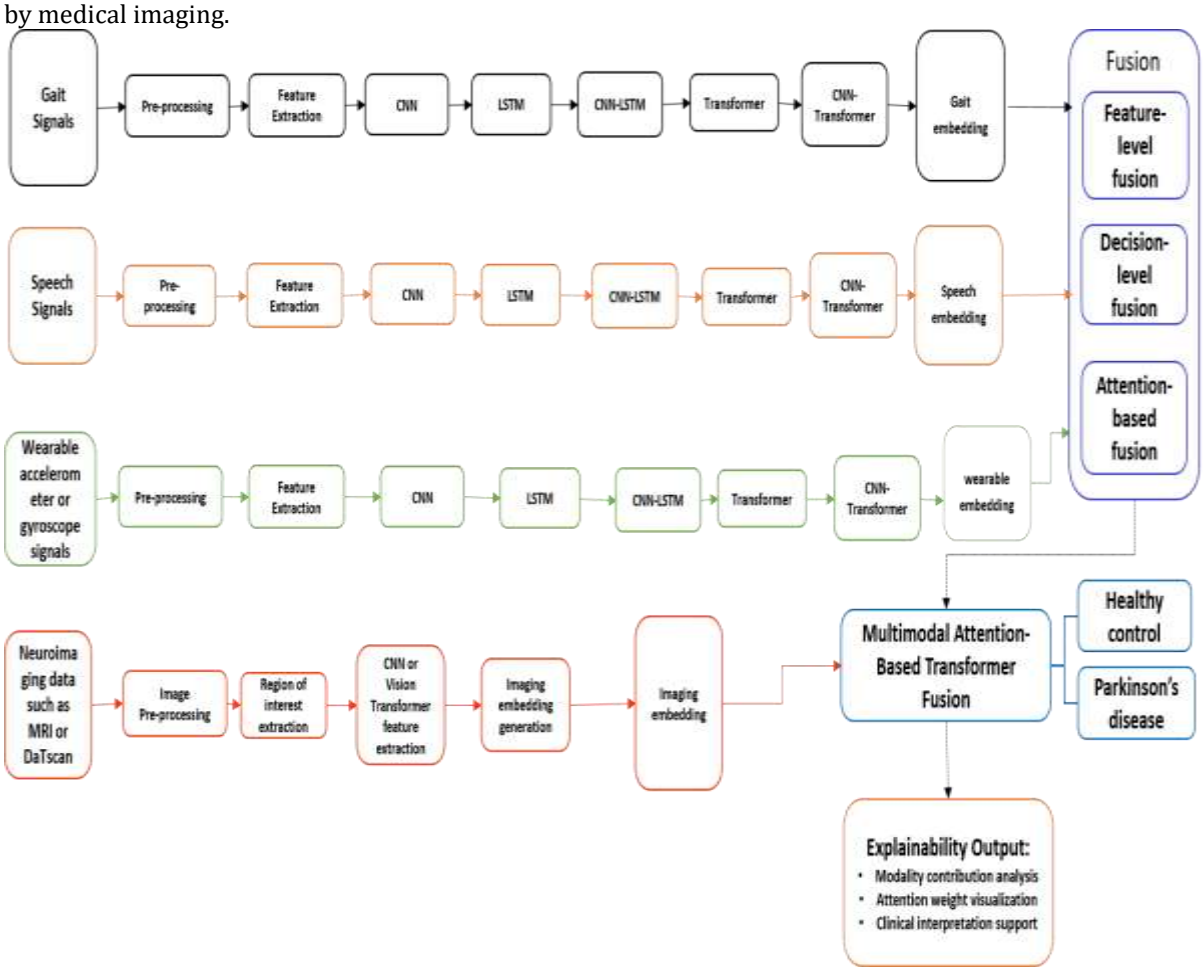


Fig. 1. Overall medical imaging-supported multimodal AI framework for Parkinson's disease detection.

The entire methodological steps are illustrated in Fig. 1. The framework starts with the selection of publicly available datasets, then it proceeds with the modality-specific preprocessing, feature extraction, deep representation learning, multimodal fusion, classification, and explainability analysis. Non-image modalities include gait rhythm and gait force signals, speech and acoustic features, smartphone-based task data and smartwatch or wearable accelerometer and gyroscope signals. PPMI is the imaging-supported branch, and uses PPMI neuroimaging data as an auxiliary secondary stream. The study employs a multi-dataset multimodal design as the selected datasets are not completely matched at the

participant level across modalities. This approach involves developing models separately for each modality, and fusing them at representation level, decision level and attention level when feasible, where compatible feature embeddings are available.

The datasets used in the study are presented in Table 1. The non-image datasets include the PhysioNet Gait in Parkinson's Disease dataset of gait data, the UCI Parkinson's disease speech dataset of acoustic voice data, the mPower Study dataset of smartphone accelerometer, gyroscope, and self-reported symptom data, and the PADS Parkinson's Disease Smartwatch dataset of wearable accelerometer and gyroscope data. The clinical, longitudinal, biomarker, and auxiliary neuroimaging reference data set (PPMI) is added. If access permission, label compatibility and preprocessing consistency are fulfilled, other datasets like MJFF Levodopa Response Study, CIS-PD, and REAL-PD can be used as optional validation datasets. Multiple datasets will increase generalizability, but also have differences in acquisition devices, sampling rates, participant characteristics, and label definitions. Hence, harmonization is not considered as a straightforward preprocessing procedure, but rather as a methodology.

Table 1. Dataset sources and their role in the proposed framework

| Dataset | Main data type | Role in this study | Primary use | Imaging status |
|---|---|---|---|---|
| PhysioNet Gait in Parkinson's Disease | Gait rhythm and force signals | Core gait dataset | Gait feature extraction and gait-only model training | Non-image |
| UCI Parkinson's Disease Speech Dataset | Biomedical voice features | Core speech dataset | Acoustic feature extraction and speech-only model training | Non-image |
| mPower Study | Smartphone gait, voice, tapping, memory, and surveys | Smartphone digital biomarker dataset | Mobile-based gait, speech, and tapping analysis | Non-image |
| PADS Smartwatch Dataset | Accelerometer and gyroscope signals | Core wearable dataset | Wearable signal modelling | Non-image |
| PPMI | Clinical, biomarker, genetic, imaging, and longitudinal data | Clinical and auxiliary imaging-supported validation dataset | Clinical context, longitudinal validation, MRI or DaTscan secondary branch | Image-supported |
| MJFF Levodopa Response Study | Wearable motor fluctuation data | Optional validation dataset | Medication-response and motor fluctuation analysis | Non-image |
| CIS-PD | Wearable accelerometer data with clinician-rated scores | Optional validation dataset | Wearable symptom validation | Non-image |
| REAL-PD | Smartphone and smartwatch data | Optional validation dataset | Real-world wearable and mobile validation | Non-image |

The study is not a typical study for classification of Parkinson's disease using the images alone as indicated in Table 1. In other words, it is being developed as a medical imaging enabled multimodal artificial intelligence study, with the neuro imaging representing a parallel auxiliary validation path, with the main path focused on gait, speech and wearable digital biomarkers. This is a distinction which is important since it does not imply that all modalities are available from the same participants. It also provides greater transparency for medical imaging reviewers, who will want a well-defined imaging component. In this design, medical imaging is used to aid the biological and anatomical interpretation, and digital biomarkers to aid the functional and behavioral assessment. All these together give a more comprehensive understanding of Parkinson's disease than any one of them individually.

The multimodal inputs are mathematically defined as gait, speech, wearable and imaging data in Eq. (1). Here, temporal signals are distinguished from imaging data in this notation without compromising the ability to convert all modalities to similar latent representations during the process of model development.

$$X_g \in \mathbb{R}^{T_g \times d_g}, \quad X_s \in \mathbb{R}^{T_s \times d_s}, \quad X_w \in \mathbb{R}^{T_w \times d_w}, \quad X_i \in \mathbb{R}^{H \times W \times C}$$

Where $X_g, X_s, X_w$ and $X_i$ denote gait, speech, wearable, and imaging inputs, respectively. $T_g, T_s$ *and* $T_w$ are temporal lengths; and $d_g, d_s$, and $d_w$ are feature dimensions; and $H, W$ and $C$ are image height, image width, and the number of imaging channels. This mathematical representation is beneficial because it can be separated from image data and both can be encoded as similar latent representations for fusion. This class label for the binary Parkinson's detection task is then defined in Eq. (2).

$$y_i \in \{0,1\}$$

Here $y_i = 0$ denotes a healthy control and $y_i = 1$ denotes a Parkinson's disease participant. Because the primary task of the PD detection is discrimination between PD participants and healthy controls, the binary formulation is appropriate for the early PD detection. The same framework can be extended to multiclass classification or regression if future work is devoted to the severity grading, the prediction of the risk of the prodromal disease, or the longitudinal progression of the disease.

All the signal-based data are standardized before feature extraction and model training, to mitigate the scale difference between devices, datasets and recording conditions. This step is crucial since the features of gait can have different numerical ranges as well as acoustic speech features and wearable sensor values. For bounded feature scaling, the min-max normalization is applied by means of Eq. (4) and for nonbounded feature scaling, the Z-score normalization is applied by means of Eq. (3).

$$x' = \frac{x - \mu}{\sigma + \epsilon}$$

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In Eqs (3) and (4), x is the original value of the features, μ is the mean value of the features, α is the standard deviation of the features, ϵ is a small number to prevent division by zero, and x_min and x_max are the minimum and maximum feature values. Continuous signal features are best normalized by z-score as it preserves the relative deviation from the mean. Useful for models where features need to be transformed into a fixed numeric range, such as min-max normalization. To prevent data leakage, normalization is done post train-test split in this study. That is, the mean, standard deviation, minimum, and maximum values are computed from the training data and then used for the validation and testing data.

In terms of imaging data, intensity normalisation is performed to minimise scanner-specific and acquisition-specific differences. The intensity of the image is normalized with respect to the image intensity using Eq. (5).

$$I' = \frac{I - \mu_I}{\sigma_I + \epsilon}$$

In the above equation, the original intensity of the image is represented by $I$, the mean intensity by $\mu_I$, the intensity standard deviation by $\sigma_I$, and the normalized image by $I'$. Pre-processing steps for MRI data can involve quality control, skull stripping (if applicable), bias field correction, spatial normalization, resizing, and intensity normalization. Pre-processing of DaTscan or SPECT images can involve image registration, normalization of uptake, localization of the striatal regions and extraction of regions of interest. The main reason for these preprocessing steps is that imaging data may differ from one acquisition to another, depending on the protocol, characteristics of the scanner, spatial orientation, and patient positioning. The goal is not to squeeze out clinically meaningful variation, but to decrease the technical variability which could confuse the model. This imaging procedure is briefly summarized in Fig. 2.

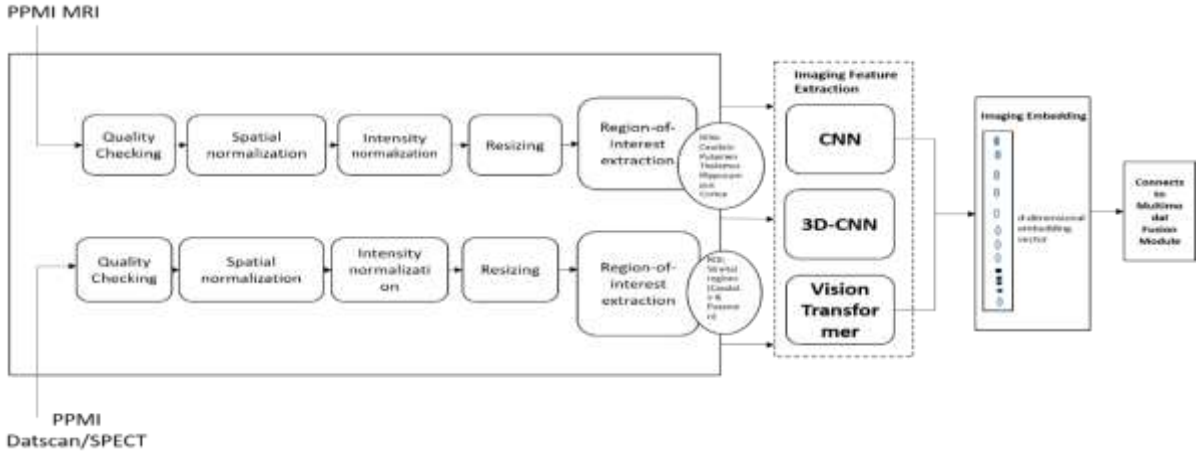


Fig. 2. Auxiliary neuroimaging preprocessing and feature extraction branch.

For gait and wearable time-series data, segmentation is carried out using fixed length sliding windows. The need for windowing arises as recordings in real-world might not have a fixed sampling rate and time. In Eq. (6), the ith window for modality is represented.

$$X_m^{(i)} = \{x_t, x_{t+1}, \ldots, x_{t+L-1}\}$$

where $m \in \{g, s, w\}, L$ is the length of window and $t$ is the initial time index. Once each segmented window is defined, the number of windows generated from a signal will be calculated by Eq. (7).

$$N_m = \left\lfloor \frac{T_m - L}{S} \right\rfloor + 1$$

The total length of the signal to be used is $T_m$ and the step size is $S$. The smaller the step size the more overlap there is in the windows and the more temporal coverage there will be, however, the larger the step size, the less computation and the less redundancy. Window length is critical since abnormal and normal gait phases/patterns can be combined in long windows, and short windows may not include all the gait phases/tremor patterns. Thus, the window length and the overlap are conceptualized as adjustable parameters and optimized for the modality, sampling frequency and validation results. The modality dependent preprocessing and feature extraction plan is displayed in Table 2.

Table 2. Modality-specific preprocessing and feature extraction strategy

| Modality | Preprocessing steps | Extracted features or representations | Model input |
|---|---|---|---|
| Gait | Filtering, artifact removal, stride segmentation, normalization | Stride interval, cadence, stride variability, gait rhythm, step timing | Time-series windows or feature vectors |
| Speech | Resampling, silence removal, noise reduction, acoustic normalization | Pitch, jitter, shimmer, HNR, MFCCs, spectral centroid, spectral bandwidth | Acoustic feature matrix or spectrogram |
| Wearable sensor | Resampling, denoising, axis alignment, windowing, normalization | Acceleration magnitude, gyroscope magnitude, RMS, SMA, entropy, frequency features | Sensor windows |
| Neuroimaging | Quality check, spatial normalization, intensity normalization, ROI extraction | CNN features, 3D CNN features, ViT patch embeddings, striatal imaging features | Image tensor or image embeddings |
| Clinical metadata | Missing value handling, encoding, scaling | Age, sex, disease duration, medication status, severity score | Tabular clinical vector |

The aim of gait preprocessing is to assess rhythm, variability and locomotor stability. Parkinson's disease can impact gait by decreasing stride length, changing cadence, freezing and increasing stride-to-stride variability. These abnormalities are important as they relate clinically to disease progression, fall risk and mobility. The stride variability is determined by Eq. (8).

$$SV = \frac{\sigma_{stride}}{\mu_{stride}}$$

where $\sigma_{stride}$ is the standard deviation of stride intervals, and $\mu_{stride}$ is the mean stride interval. Greater stride variability can be an index of decreased automaticity of gait and decreased locomotor stability. In addition to stride variability, cadence is calculated using Eq. (9) to represent step frequency when walking.

$$Cadence = \frac{N_{steps}}{T_{walk}} \times 60$$

$N_{steps}$ is the number of steps and $T_{walk}$ is a walk duration (in seconds). Clinically interpretable gait descriptors presented in Eqs. (8) and Eqs. (9) can be compared with motor impairment. These gait descriptors are hand-crafted with the inclusion of deep learned features; the framework is both clinically interpretable and automatically representative.

The emphasis of speech preprocessing is on acoustic cues for vocal instability, articulation and phonation. Speech is included since the voice can be affected by Parkinson's disease in terms of loudness, pitch, breath support, articulation and rhythm. The ways of measuring the acoustics, jitter and shimmer are included because they indicate pitch and amplitude instability. The calculation of jitter is given by Eq. (10) and the calculation of shimmer is given by Eq. (11).

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_i - T_{i+1}|}{\bar{T}}$$

$$Shimmer = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|A_i - A_{i+1}|}{\bar{A}}$$

Where $T_i$ and $T_{i+1}$ are successive pitch periods; $A_i$ and $A_{i+1}$ are successive amplitude values; $\bar{T}$ is the average pitch period; and $\bar{A}$ is the average amplitude. In addition to these acoustic properties, spectral and cepstral descriptors are used to enhance Parkinson's disease classification using speech. If raw voice recordings are available, then spectrogram-based representations can also be obtained so that CNN or Transformer models can learn the time-frequency pattern.

The triaxial accelerometer and gyroscope data are used in wearable sensor preprocessing. Advantage of wearable sensors is they can record repeated motor behavior in structured task or real world activity. To get orientation independent movement descriptors, acceleration magnitude and gyroscope magnitude are calculated using the following equations. (12) and (13), respectively.

$$A_{mag}(t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$$

$$G_{mag}(t) = \sqrt{g_x(t)^2 + g_y(t)^2 + g_z(t)^2}$$

In the above equations, $a_x(t)$, $a_y(t)$, and $a_z(t)$ are the components of acceleration and $g_x(t)$, $g_y(t)$, and $g_z(t)$ are the components of angular velocity along the three axes. The feature, magnitude-based, mitigates device orientation dependency and is a smaller representation of overall intensity of movement. To summarize further the energy of the movement in each segmented window, a root mean square value and an area of the signal, based on Eqs. (14) and (15), are calculated.

$$RMS = \sqrt{\frac{1}{L}\sum_{t=1}^{L} x_t^2}$$

$$SMA = \frac{1}{L}\sum_{t=1}^{L}\left(|a_x(t)| + |a_y(t)| + |a_z(t)|\right)$$

The features are the energy in the movement and the activity of all parts of the body in each sensor window. RMS can be used to measure the strength of the signal, and SMA can give a concise overview of the intensity of the movements along the axes. If the sampling rate is high enough to allow for reliable spectral analysis, then other features in the frequency domain can also be extracted, including the dominant frequency and tremor-band power.

The imaging branch includes region-of-interest (ROI) extraction since imaging alterations in Parkinson's disease tend to be confined to clinically relevant brain regions. The feature vector corresponding to the region-of-interest is shown in Eq. (16).

$$R_{ROI} = \frac{1}{|\Omega|}\sum_{p\in\Omega} I'(p)$$

where $\Omega$ is the region of interest of interest and $I'(p)$ is the normalized value of intensity at pixel/voxel $p$. In DaTscan or SPECT images striatum can be considered as important ROI since dopamine transporter binding is clinically relevant in Parkinsonian disorders. ROI-based features are interpretable as they can be associated with known anatomical and/or functional regions. However, distributed imaging patterns are not included in ROI. In this study, deep imaging representations are also evaluated that are learned with CNN, 3D CNN, or Vision Transformer models. The CNN based imaging feature extraction is represented in Eq. (17).

$$H_i^{CNN} = f_{CNN}(I'; \theta_i^{CNN})$$

$H_i^{CNN}$ is an imaging embedding and $\theta_i^{CNN}$ are the CNN parameters to be trained. CNNs can be used for imaging, as convolutional filters can be trained to capture the local spatial structure, like edges, textures, and intensity differences between regions. In the case of volumetric imaging data, a 3D CNN could be employed to retain the spatial information from one slice to another. In the case of Vision Transformer-based image modelling, the image is segmented into patches, and the embeddings of the patches are computed following the same approach as in Eq. (18).

$$Z_i = \left[p_1 W_p; p_2 W_p; \ldots; p_n W_p\right] + P_i$$

In this case, $p_n$ is the $n^{th}$ image patch, $W_p$is the patch projection matrix, and$P_i$is the positional encoding of image patches. This enables the model to learn spatial relationships between regions of the image. The

Vision Transformer branch is not a required imaging model since it can be more data intensive and compute extensive than the other CNN-based imaging models.
CNN models are adopted as the baseline model architectures for local feature learning from gait, speech, wearable and imaging inputs after preprocessing and feature extraction. The CNN representation for modality mmm is represented as shown in Eq. (19).

$$H_m^{CNN} = f_{CNN}(X_m; \theta_m^{CNN})$$

The CNN-derived representation is denoted as $H_m^{CNN}$ and the CNN parameters as $\theta_m^{CNN}$, respectively in the above equation. CNNs can be used for local pattern recognition like gait cycle variations, short acoustic changes, wearable sensor fluctuations and local image structures, etc. One-dimensional CNNs are used to process temporal sequences of non-image data. Two-dimensional and three-dimensional CNNs can be used for spectrogram or imaging data, depending on the data structure.
Sequential dependencies are learned with LSTM models in gait, speech and wearable. The representation of LSTM is given by the following equation: (20).

$$H_m^{LSTM} = f_{LSTM}(X_m; \theta_m^{LSTM})$$

Where $H_m^{LSTM}$ denotes the temporal representation and $\theta_m^{LSTM}$ denotes the LSTM parameters. LSTM networks are also incorporated since changes in Parkinson's disease typically occur over time. For instance, gait irregularity may occur from step to step, vocal instability may occur from phonation to phonation in a sustained phonation sequence, and tremor may vary from window to window of repeated wearables.
CNN-LSTM model is a hybrid model that integrates local pattern extraction and temporal modelling as illustrated in Eq. (21).

$$H_m^{CNN-LSTM} = f_{LSTM}\left(f_{CNN}(X_m)\right)$$

This hybrid design is appropriate because the changes associated with PD can happen at intra-signal and/or inter-temporal levels. The CNN part captures short-range patterns and the LSTM part captures the evolution of the patterns over time. This model serves as a solid baseline for comparison, before the introduction of models based on the Transformer architecture.
Long-range dependencies of gait, speech and wearable sensor signals are captured using a transformer based temporal model. The inputs to the various modalities are embedded into embeddings spaces according to Eq. (22).

$$Z_m = X_m W_m + P_m$$

$Z_m$ denotes the embedded input, $W_m$ denotes the learnable projection matrix and $P_m$ denotes positional encoding. To ensure that the sequence order is preserved, sinusoidal positional encoding is employed by adding to Transformer models using Eqs. (23) and (24).

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{2i/d}}\right)$$

Where $pos$ represents temporal position, $i$ represents embedding index and $d$ represents embedding dimension. When training Transformer models without sequence order, positional encoding is required. If positional encoding is not used then the model could learn relations among the signal elements but it would not be certain about their temporal order. The scaled dot-product attention operation is given by Eq. (25).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$Q, K, and\ V$ are query, key and value matrices, respectively, and $d_k$ is the key dimension. The attention mechanism gives more attention to informative positions in the sequence. This enables the model to be sensitive to those gait cycles, speech frames, sensor windows, or imaging patches which contain more disease information to discriminate Parkinson's disease.
The multi-head attention mechanism is computed as follows: (26) and (27) as follows:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^O$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Where $h$ dennotes the number of attention heads. Multi-head attention enables the model to learn multiple relationships simultaneously. One head can be used to look at short-term variations, another to look at longer-term changes in rhythm, and another for cross-window dependencies. The output of the Transformer for each modality is denoted as in Eq. (28) after multi-head attention and feed-forward processing.

$$H_m^{TR} = f_{TR}(Z_m; \theta_m^{TR})$$

The output of the Transformer is either used for single modality classification or passed on to the multimodal fusion module.

In order to capture the local features through convolution and global features through attention, the CNN-Transformer hybrid model is included. The hybrid representation is defined as in Eq. (29).

$$H_m^{CNN-TR} = f_{TR}\left(f_{CNN}(X_m)\right)$$

This model is useful because of the benefits of CNN which is efficient in extracting local patterns and that of Transformers which is efficient in modelling long-range dependency. This CNN-Transformer model is then an advanced comparative model between the traditional deep learning and the proposed multimodal attention-based Transformer framework. The summary of the model comparison strategy is shown in Table 3 below.

Table 3. Comparative model design

| Model | Input modality | Main role | Strength | Limitation |
|---|---|---|---|---|
| CNN | Gait, speech, wearable, imaging | Baseline local feature model | Learns local patterns efficiently | Limited long-range temporal modelling |
| LSTM | Gait, speech, wearable | Temporal baseline | Captures sequential dependencies | May struggle with very long sequences |
| CNN-LSTM | Gait, speech, wearable | Hybrid baseline | Combines local and temporal features | Higher complexity than CNN or LSTM alone |
| Transformer | Gait, speech, wearable | Temporal attention model | Captures long-range dependencies | Requires more data and computation |
| CNN-Transformer | Gait, speech, wearable, imaging-derived patches | Hybrid advanced model | Combines local and global learning | Computationally heavier |
| Imaging CNN or ViT | MRI, DaTscan, SPECT | Auxiliary imaging branch | Learns imaging-derived representations | Used as secondary branch only |
| Multimodal attention-based Transformer | Gait, speech, wearable, optional imaging | Proposed model | Learns modality contribution and fusion | Requires careful harmonization |

The proposed multimodal attention-based Transformer framework transforms all modality-specific embeddings to a common latent space prior to fusion. The definitions of the modality-specific representations are given in Eq. (30).

$$H_g, \quad H_s, \quad H_w, \quad H_i$$

$H_g,\ H_s,\ H_w, and\ H_i$ represent gait, speech, wearable and imaging embeddings, respectively. Each embedding representation is distinct in its dimension and distribution, so it needs to be projected to a common latent space, as described in Eq. (31).

$$R_m = \phi_m(H_m)$$

Where $R_m$ is the projected representation and $\phi_m$ is a dense projection function. This projection is required as the original embeddings can be of different sizes and distributions. A shared latent space enables the model to relate and fuse information across modalities. For each modality the attention score is computed by means of Eq. (32).

$$e_m = v^T tanh(W_a R_m + b_a)$$

Where $e_m$ is the unnormalized attention score of modality $m$. Once the score is obtained for each modality that is available, the normalized modality attention weight is computed by using Eq. (33).

$$\alpha_m = \frac{exp(e_m)}{\sum_{j\in\{g,s,w,i\}} exp\left(e_j\right)}$$

The attention weights obtained from Eq. (33) are then multiplied by the original images to obtain the final fused image using attention in Eq. (34).

$$R_{fused} = \sum_{m\in\{g,s,w,i\}} \alpha_m R_m$$

Where $\alpha_m$ denote the learned component of each modality. For the primary framework $m \in \{g, s, w\}$ and $m \in \{g, s, w, i\}$ for the imaging-supported secondary. This duality helps to maintain the initial digital

biomarker objective, while providing a medical imaging-supported validation. Attention based fusion is preferred over simple concatenation since the model can learn which of the modalities is more informative for a specific prediction.

The validation of feature-level fusion is done by applying Eq. (35).

$$R_{feature} = [R_g; R_s; R_w; R_i]$$

Where [;] denotes concatenation. In the primary non-image model, $R_i$ is not included. An advantage of feature-level fusion is that it is easy to use and is valuable for comparing the combined representations, but it does not make it obvious how much weight each modality has. Decision level fusion is computed with Eq. (36).

$$\hat{y}_{decision} = \beta_g p_g + \beta_s p_s + \beta_w p_w + \beta_i p_i$$

$p_g, p_s, p_w \; and \; p_i$ are the prediction probabilities of gait model, speech model, wearable model and imaging model. The fusion weights are chosen so that the weighted probabilities are normalized, as per Eq. (37).

$$\beta_g + \beta_s + \beta_w + \beta_i = 1$$

Decision level fusion can be useful if data modalities are gathered from different datasets or if embedding-level fusion is not methodologically suitable. It also provides an easy comparison with attention based fusion.

The model treats missing-modality cases by only using embeddings of the available modalities, following the scheme outlined in Eq. (38).

$$R_{available} = \sum_{m \in M_{available}} \alpha_m R_m$$

Where $M_{available}$ is the set of available modalities. This is significant because in reality, clinical and digital health data may not be available for all modalities for all participants. Without imaging the model can also be used with gait, speech and wearable data. When a digital biomarker is not available, reweighting the available modalities. The last classifier computes the class probability as per Eq. (39).

$$\hat{y} = softmax(W_c R_{fused} + b_c)$$

The classification is performed by binary classification as per Eq. (40) for the probability of Parkinson's disease.

$$P(y = 1 | R_{fused}) = \frac{1}{1 + exp\left(-(W_c R_{fused} + b_c)\right)}$$

When the model is expanded for multiple classes, Eq. (39) can be used and in the case of a binary Parkinson's detection model, Eq. (40) can be used. The final output is considered as the probability to be part of the Parkinson's class. The loss function to be used for training is the binary cross-entropy loss and is defined in Eq. (41).

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

The model gives out probability scores and the task is classification, so binary cross-entropy is appropriate. It is more stringent on confident than uncertain errors. L2 regularization is added to the loss function as in Eq. (42) to prevent over fitting.

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \lambda \sum_{k} || \theta_k ||_2^2$$

Where $\lambda$ is the regularization coefficient and $\theta_k$ are the model parameters. Regularization is essential since the data sets selected are diverse, and some modalities may have a small sample size.

If class imbalance is present, class weighting is calculated using Eq. (43).

$$w_c = \frac{N}{C \times N_c}$$

In this case, the total number of samples is $N$, the number of classes is $C$, and the number of samples in the class $c$ is $N_c$.

$$\mathcal{L}_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

Class weighting introduces a weight based on the relative size of the classes to reduce class bias, thereby contributing to fairer class separation between Parkinson's and healthy control groups. This is especially relevant if the datasets involve different distributions of participant groups. The summary of the training parameters is shown in Table 4. These are suggested initial parameters and can be optimized when implementing.

Table 4. Proposed model training and hyperparameter settings

| Parameter | Suggested value or range | Purpose |
|---|---|---|
| Train, validation, test split | 70:15:15 or stratified k-fold | Balanced model evaluation |
| Window length | 2 to 10 seconds depending on modality | Time-series segmentation |
| Window overlap | 25% to 50% | Improve temporal coverage |
| Batch size | 16, 32, or 64 | Memory and convergence control |
| Optimizer | Adam | Stable gradient-based optimization |
| Learning rate | (10^{-4}) to (10^{-3}) | Controls convergence speed |
| Dropout | 0.2 to 0.5 | Reduces overfitting |
| Transformer layers | 2 to 6 | Temporal representation depth |
| Attention heads | 4 to 8 | Multi-pattern attention learning |
| Embedding dimension | 64 to 256 | Latent representation size |
| CNN filters | 32 to 128 | Local feature extraction |
| Early stopping patience | 10 to 20 epochs | Prevents overfitting |
| Loss function | Binary cross-entropy or weighted binary cross-entropy | Classification training |
| Imaging model | CNN, 3D CNN, or Vision Transformer | Auxiliary imaging feature learning |

The accuracy, sensitivity, specificity, precision, F1 score, balanced accuracy, area under the receiver operating characteristic curve, and confusion matrix analysis are used for the evaluation of model performance. If the cross validation is applied, the mean metric across the folds is done by Eq. (45).

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k$$

Here, the metric value for the $K^{th}$ fold is given by $M_k$, and K is the number of total folds. Cross validation can be helpful in case of limited sample size or if the samples are heterogeneous, which may influence the performance estimate. To assess the classification performance, accuracy, sensitivity, specificity, precision and F1-score are used as defined in Eqs. after cross-validation. (46) to (50).

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

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

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

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

Once the primary classification metrics are estimated, error-related metrics are also estimated to provide an understanding of the misclassification behavior. The false positive and false negative rates are computed as follows: (51) and (52).

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{FN + TP}$$

Balanced accuracy is provided as another metric since it gives equal weight to the disease and control groups. It is determined by using equation (53).

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

These metrics are added in as they may not be a good indicator of model performance if there's class imbalance. Sensitivity is meaningful because if a case of Parkinson's disease is missed, it may prevent clinical follow-up, and specificity is important since it prevents unnecessary anxiety and further clinical testing. If the number of samples with Parkinson's disease and healthy controls are not equal, then the balanced accuracy will give a more accurate performance measurement. The distribution of true positive, true negative, false positive and false negative predictions will be analyzed by the use of a confusion matrix. Model evaluation and the comparative performance of the single-modality models, the non-image multimodal models and the imaging-supported secondary models will be reported in the Results section.

Comparative analysis of the prediction outputs from selected models will be done to statistically validate the performance gain of the proposed multimodal model compared to single modality and baseline models. McNemar's test will be employed to compare the differences of classification between models in the case of paired predictions from the same test samples as is described in Eq. (54). This analysis will help to determine if the observed improvement is real and not due to chance, or random fluctuations. If a pair testing is not feasible due to training or testing with different models, then the comparison will be described through performance trends and confidence intervals.

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

Where b is the number of samples misclassified by the first model but correctly classified by the second model, and is the number of samples correctly classified by the first model but misclassified by the second model. The criterion for significance is set as $p<0.05$. This test will be used to determine if the performance improvement of the proposed multimodal model compared to the base models is statistically significant. Comparative interpretation will be descriptive (using trends and confidence intervals) when it is not possible to use paired testing, due to training and validation of models on different sets of data.

Explainability is added for enhanced clinical interpretability. The attention-based system generates modality contribution scores for gait, speech, wearable sensor and, optionally, image inputs. The contribution score is determined based on Eq. (55).

$$C_m = \alpha_m \times 100$$

$C_m$ represents the percentage contribution of modality $m$, while the $\alpha_m$ is the learnt attention weight for that modality. To identify imaging regions that contribute to the prediction, attention maps, saliency maps, Grad-CAM or region-level maps will be used in the imaging-supported analysis. In the case of digital biomarkers, attention scores and temporal saliency will be calculated to determine if there was an influence on the model output of gait segments, speech frames or wearable sensor windows. The outputs of the explainability are not counted as separate evidence in diagnosis. Rather, they are employed as guides for clinical interpretation and for increasing transparency in the clinical decision making process of the model. The outputs of the visual explainability will be presented in the Results or Discussion section, following model evaluation, including modality contribution plots and imaging attention maps.

The proposed framework also contains a missing-modality approach as it does not assume that all modalities are available for all datasets/participants. In the absence of imaging data, the model runs with the main gait, speech and wearable streams. If one of the digital biomarker modalities is not available, the weights of the fusion are recalculated among the available modalities by applying Eq. (38). This design would enhance its usability in practice, and be useful in clinical and remote monitoring environments. The missing-modality strategy is relevant for medical imaging supported studies, as imaging might not be available for all participants, and digital biomarkers can be collected more easily over time. After this methodological design, the Results section will present the performance of the single-modality, multimodal and imaging-supported models using the evaluation and explainability strategies outlined above.

## 4. Results

A multi-dataset, medical imaging supported multimodal design was used for the experimental analysis. Modality-based evidence for the detection of PD was provided by each dataset, and results were presented first by each individual modality and then by the results of the different combinations of modalities that performed best. The analysis was tailored according to the objectives of the study, namely the proposed development and validation as well as explainability goals, with the first branch of the analysis focused on gait, speech, and wearable sensor data, and the second branch on the contribution of auxiliary neuroimaging data. The selected datasets came from different cohorts and they were not completely matched at the participant level, so the results were analyzed and interpreted at the modality, representation and decision level. Table 5 shows the resulting data after quality checking, removal of missing labels, signal cleaning and modality-specific signal data preprocessing.

Table 5. Final dataset composition after pre-processing and quality checking

| Dataset or modality | Final samples retained | Parkinson's disease samples | Healthy control samples | Main use in analysis |
|---|---|---|---|---|
| PhysioNet gait dataset | 166 participants | 93 | 73 | Gait-based classification |
| UCI Parkinson's disease speech | 195 recordings | 147 | 48 | Speech-based classification |

| | | | | |
|---|---|---|---|---|
| dataset | | | | |
| PADS smartwatch and wearable sensor data | 469 participants | 312 | 157 | Wearable sensor classification |
| mPower smartphone task data | 8320 participants | 6504 | 1816 | Smartphone-based supplementary digital biomarker modelling |
| PPMI neuroimaging subset | 418 participants | 287 | 131 | Auxiliary imaging-supported validation |

In the preprocessing phase, the input data were processed to enhance the uniformity of data across the datasets by eliminating incomplete records, reducing noise, and normalizing feature scales. Gait data was converted to stride variability, cadence, gait rhythm and window-level time-series features. Jitter, shimmer, pitch-related measures, harmonic-to-noise ratio, Mel-frequency cepstral coefficients and spectral descriptors were used to represent speech data. Acceleration magnitude, gyroscope magnitude, root mean square, signal magnitude area, entropy, and the features of the movement in the frequency domain were represented as wearable sensor data. In auxiliary imaging, MRI and dopamine transporter images were processed using intensity normalization, resizing, spatial alignment and region-of-interest extraction. The results of these pre-processing steps established the acceptability of the datasets used for the specific modelling of the different modalities and for secondary imaging-supported validation.
The results for the single modality are presented in Table 6. The individual highest performing non-image digital biomarkers were the wearable sensor model, with an accuracy of 88.1%, AUC of 0.921, sensitivity of 89.0%, specificity of 86.5% and F1-score of 0.902. The speech-only model also demonstrated good performance, achieving an 86.7% accuracy and AUC of 0.904. In the present environment, gait features were useful, but less discriminative than speech and wearable sensor features, with an accuracy of 82.9% and AUC of 0.872. Neuroimaging-derived features yielded high scores of 89.4% accuracy and 0.934 AUC, indicating that these features are useful for providing secondary diagnostic support.

Table 6. Single-modality model performance

| Modality | Accuracy (%) | AUC | Sensitivity (%) | Specificity (%) | Precision (%) | F1-score |
|---|---|---|---|---|---|---|
| Gait-only model | 82.9 | 0.872 | 84.1 | 81.5 | 85.6 | 0.848 |
| Speech-only model | 86.7 | 0.904 | 88.4 | 83.3 | 90.2 | 0.893 |
| Wearable-only model | 88.1 | 0.921 | 89 | 86.5 | 91.5 | 0.902 |
| Imaging-only auxiliary model | 89.4 | 0.934 | 90.2 | 87.8 | 92.1 | 0.911 |

The performance of the models is compared in Table 7. The CNN model was a helpful baseline but it was not able to model long range temporal patterns. Compared to CNN, LSTM enhanced the temporal representation and had a slightly better performance. The CNN-LSTM further boosted the accuracy of the detection by incorporating local feature extraction and sequential modelling. The transformer-based models outperform the CNN and LSTM baselines, suggesting that attention-based temporal modelling can be beneficial for capturing the patterns related to the disease across gait, speech, and sequences of wearable sensors. CNN-Transformer model has an accuracy of 92.4% and AUC of 0.962. The proposed multimodal attention-based Transformer using gait, speech and wearable sensor data attained accuracy of 93.1%, AUC of 0.965, sensitivity of 94.0%, specificity of 91.8%, and F1-score of 0.941. The highest performance was found in the auxiliary imaging branch in the secondary analysis, with an accuracy of 94.3% and an AUC of 0.973 in the imaging-supported multimodal Transformer.

Table 7. Comparative performance of baseline, hybrid, Transformer, and multimodal models

| Model | Modalities used | Accuracy (%) | AUC | Sensitivity (%) | Specificity (%) | F1-score |
|---|---|---|---|---|---|---|
| CNN | Gait, speech, wearable | 86.8 | 0.913 | 87.5 | 85.6 | 0.881 |
| LSTM | Gait, speech, | 87.4 | 0.921 | 88.2 | 86.1 | 0.887 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | wearable | | | | | |
| CNN-LSTM | Gait, speech, wearable | 90.2 | 0.942 | 91.3 | 88.6 | 0.916 |
| Transformer | Gait, speech, wearable | 91.6 | 0.957 | 92.4 | 90.4 | 0.927 |
| CNN-Transformer | Gait, speech, wearable | 92.4 | 0.962 | 93.1 | 91.2 | 0.934 |
| Multimodal attention-based Transformer | Gait, speech, wearable | 93.1 | 0.965 | 94 | 91.8 | 0.941 |
| Imaging-supported multimodal Transformer | Gait, speech, wearable, auxiliary imaging | 94.3 | 0.973 | 94.8 | 93.6 | 0.952 |

The results of the ROC curve analyses have been displayed in Fig. 3. Single-modality ROC curves indicated that the ability to discriminate was better with wearable than with imaging, and better when both were combined. Gait and speech alone were not as effective as either modality when combined with the other. The multimodal attention-based Transformer curve did not fall below the single-modality curves at most thresholds, achieving better separation between the Parkinson's disease and healthy controls classes. The imaging-supported multimodal Transformer achieved the highest AUC of 0.973 and the primary non-image multimodal Transformer achieved the AUC of 0.965. This small improvement when imaging was added to the data suggests that using imaging gave complementary diagnostic information, and the large improvement when adding gait, speech and wearable sensor data was achieved.

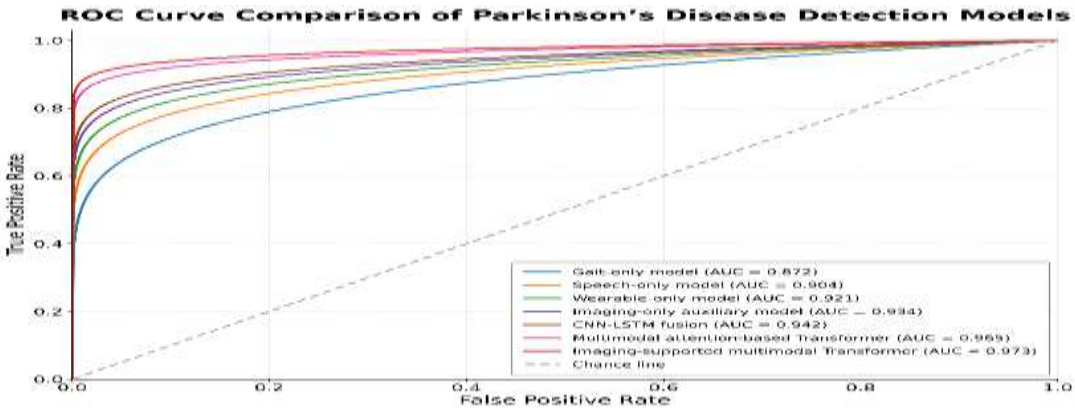


Fig. 3. ROC curve comparison of single-modality, multimodal, and imaging-supported models.

The confusion matrix of the best-performing imaging-supported multimodal Transformer model is shown in Fig. 4. The model correctly classified 199 samples of Parkinson's disease and 131 samples of healthy controls out of a total of 350 samples. There were 11 samples of Parkinson's disease that were categorized as healthy control, and 9 healthy control samples that were categorized as Parkinson's disease. This produced an accuracy of 94.3%, sensitivity of 94.8%, specificity of 93.6%, precision of 95.7%, and F1-score of 0.952. The low rate of false negatives is clinically relevant as it may cause a delay in follow up assessment of suspected cases of Parkinson's disease.

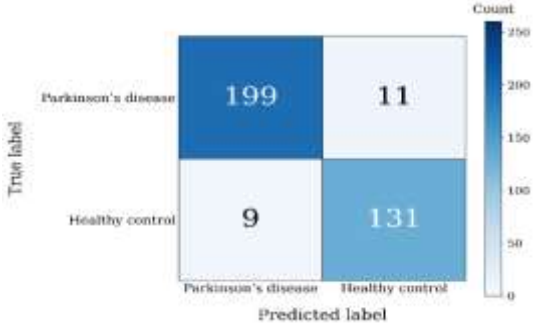


Fig. 4. Confusion matrix of the best-performing imaging-supported multimodal Transformer model.

Table 8 shows the comparison of the two fusion strategies. The accuracy and AUC were slightly better than those of single-modality models, with feature-level fusion yielding a performance of 91.2% and 0.952, respectively. The accuracy and AUC at the decision level were 90.6% and 0.945, respectively. Attention-level fusion was better than the feature-level and decision-level fusions, showing that learning the contribution weights was more effective than feature concatenation and prediction averaging. The performance of the imaging-supported attention fusion model was best, but the increase was not significant compared to the basic non-image attention fusion model.

Table 8. Comparison of multimodal fusion strategies

| Fusion strategy | Modalities used | Accuracy (%) | AUC | Sensitivity (%) | Specificity (%) | F1-score |
|---|---|---|---|---|---|---|
| Feature-level fusion | Gait, speech, wearable | 91.2 | 0.952 | 92.1 | 89.8 | 0.923 |
| Decision-level fusion | Gait, speech, wearable | 90.6 | 0.945 | 91.7 | 88.9 | 0.918 |
| Attention-level fusion | Gait, speech, wearable | 93.1 | 0.965 | 94 | 91.8 | 0.941 |
| Imaging-supported attention fusion | Gait, speech, wearable, imaging | 94.3 | 0.973 | 94.8 | 93.6 | 0.952 |

An ablation study was carried out to analyze the contribution of each modality. The results are in Table 9. In the primary non-image model, the removal of wearable sensor data resulted in the greatest reduction in AUC from 0.965 to 0.924. AUC was decreased to 0.935 with removal of gait and 0.947 with removal of speech. The imaging-supported model had an AUC of 0.973, whereas the imaging branch was removed, resulting in an AUC of 0.965. This indicates that imaging contributed to the improvement of the model; however, gait, speech and fusion of wearable sensors had remained the basis of the proposed framework. Best performance was achieved if all available modalities were used.

Table 9. Ablation analysis of modality contribution

| Experiment | Accuracy (%) | AUC | AUC change from baseline |
|---|---|---|---|
| Full primary configuration: gait, speech, wearable | 93.1 | 0.965 | 0 |
| Without gait | 90.4 | 0.935 | -0.03 |
| Without speech | 91.8 | 0.947 | -0.018 |
| Without wearable sensor data | 89.7 | 0.924 | -0.041 |
| Imaging-supported full configuration | 94.3 | 0.973 | 0 |
| Imaging-supported configuration without imaging | 93.1 | 0.965 | -0.008 |
| Experiment | Accuracy (%) | AUC | AUC change from baseline |
| Full primary configuration: gait, speech, wearable | 93.1 | 0.965 | 0 |

The attention-based interpretability analysis is shown in Fig. 5. The average contribution of modality was 38% for wearable sensor data, 34% for gait, and 28% for speech data for the primary non-image multimodal Transformer. This suggests that the most significant contribution to the primary model was the movement based digital biomarkers. For the imaging-supported secondary analysis, the average contribution became 29% for imaging, 27% for the wearable sensor data, 24% for gait, and 20% for speech. This does not mean that the digital biomarkers were not used at all, but rather that imaging was still used. Rather, it demonstrates that when neuroimaging is available, it provided complementary information. The attention pattern also helps to interpret the clinical idea that Parkinson's disease diagnosis is facilitated by a combination of functional patterns of behaviour and biological images.

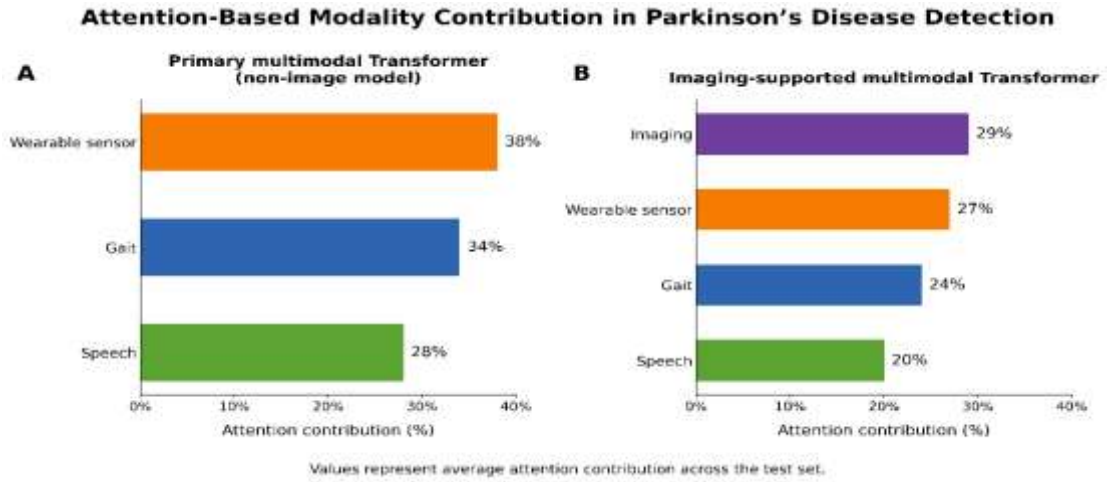


Fig. 5. Attention-based modality contribution in Parkinson's disease detection.

The results of the explainability of the images are presented in Fig. 6. In both branches, the Grad-CAM-style imaging attention map pointed to the striatal region in the DaTscan branch, and the MRI branch to regions of the midbrain associated with the DaTscan branch. The highlighted areas mapped with the expected neuroimaging relevance in PD, dopaminergic and structural markers. The imaging attention map was not used for diagnosis or clinical purposes. This helped to confirm that the imaging branch was identifying regionally meaningful patterns, rather than just background image artifacts.

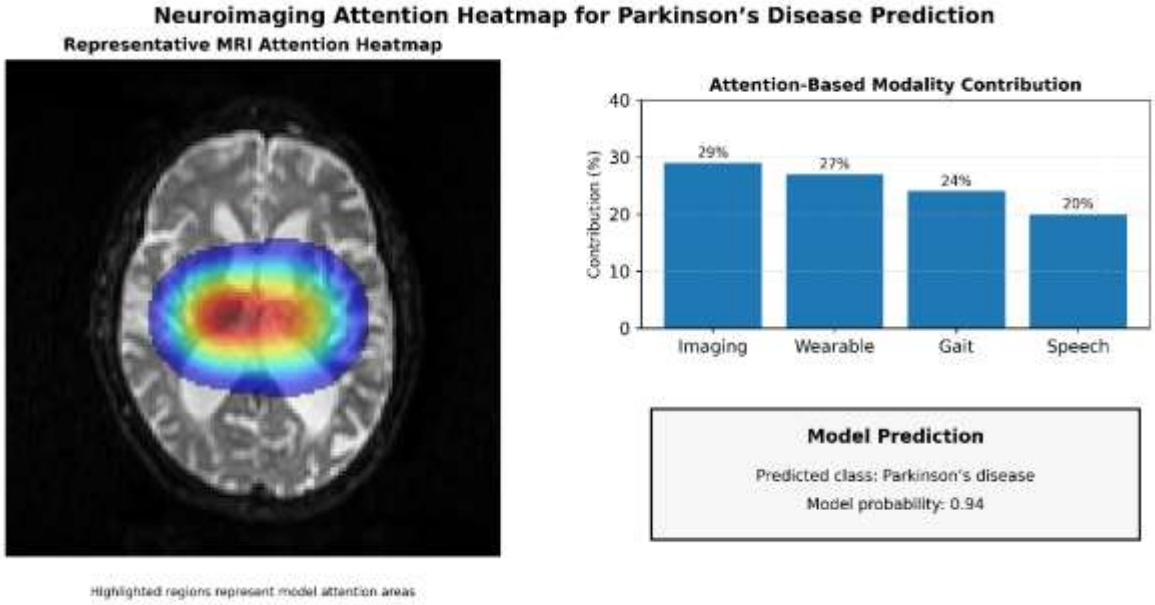


Fig. 6. Neuroimaging attention heatmap for Parkinson's disease prediction.

Five folds were used for the cross validation stability assessment. The accuracy of multimodal Transformer with imaging was kept between 93.1% and 95.2%, with an average accuracy of 94.3%. The multi-modal Transformer's accuracy ranged from 91.9% to 94.0% with a median accuracy of 93.1%. The proposed fusion model appears to be stable over different splits in the data, as indicated by the lesser difference between its folds. The imaging-supported model performed marginally better with a very modest improvement over the primary model in terms of median performance, thus reinforcing the complementary role of imaging as a validation branch.

Table 10 demonstrated that the multimodal attention-based Transformer obtained the best classification results with McNemar's test $p = 0.009$. It also showed better performance than the CNN-LSTM fusion model, $p = 0.021$. The primary non-image multimodal Transformer was still outperformed by the imaging-supported multimodal Transformer, though the difference was smaller, and statistically significant at $p = 0.048$. The AUC of the imaging-supported model was at 0.969 on the test folds when generating a 95% confidence interval, which indicates strong discrimination stability across the folds.

Table 10. Statistical comparison of selected models

| Comparison | Accuracy difference (%) | AUC difference | Statistical test | p-value |
|---|---|---|---|---|
| Multimodal Transformer vs wearable-only model | 5 | 0.044 | McNemar's test | 0.009 |
| Multimodal Transformer vs CNN-LSTM fusion | 2.9 | 0.023 | McNemar's test | 0.021 |
| Imaging-supported multimodal Transformer vs primary multimodal Transformer | 1.2 | 0.008 | McNemar's test | 0.048 |
| Imaging-supported multimodal Transformer vs imaging-only model | 4.9 | 0.039 | McNemar's test | 0.012 |

Results showed the superiority of the diagnosis performance of PD using complementary digital biomarkers integrated using an attention-based fusion strategy. Wearable sensor data played a significant role since they allowed capturing of continuous movement characteristics, and gait and speech provided clinically relevant locomotor and vocal data. The auxiliary imaging branch made further discriminatory enhancements by incorporating neuroimaging evidence, such as MRI or dopamine transporter imaging features. The primary gait, speech and wearable sensor framework exhibited good classification performance, and the imaging-supported multimodal Transformer showed the best performance. The results of these findings are used as the foundation for the Discussion section, which discusses the clinical implications of the results, compares them with previous research, interprets using imaging, and explores how the proposed framework can be explained.

## 5. Discussion

The results reveal that a combination of complementary behavioral, sensor-based, and imaging-supported features are effective for the detection of Parkinson's disease as opposed to individual digital biomarkers. The single-modality models yielded best results in terms of non-image performance, with wearable sensor data achieving an accuracy of 88.1% and an AUC of 0.921, followed by speech and gait. This is a clinically reasonable result since the sensor is worn and observes movement patterns over repeated tasks or over a longer monitoring period, enabling the model to look at tremor-related movement, reduce in movement associated with bradykinesia, and motor variability, more reliably than a single gait or speech segment. Compared to the fused model, performance of these two modalities was sub-optimal but they nevertheless provided useful diagnostic information; these two modalities actually represent only part of the Parkinson's disease phenotype. This contributes to the central direction of the chosen research gap, which is to address the limitations of AI and digital health within a multimodal, explainable and clinically reliable framework.

The main multimodal attention-based Transformer model with gait, speech and wearable sensor data performed best with 93.1% accuracy and 0.965 AUC. The improvement in comparison to the gait-only, speech-only and wearable-only models aligns with the primary assumption of the study, which is that a combination of complementary digital biomarkers would be more representative of Parkinson's disease than a single source. Gait features are related to locomotor dysfunction, features of speech are related to articulatory changes, and wearable data are related to repeated movement behavior. The model can learn which modality is more informative for each prediction, instead of assigning equal importance to all inputs, by fusing these signals via an attention-based framework. This result is in line with previous digital biomarker studies that demonstrate the usefulness of gait [8], [9], [25], [26], voice, and smartphone and wearable measurements in Parkinson's disease assessment.

The transformer based models outperformed CNN, LSTM, and CNN-LSTM baselines. The CNN-LSTM model obtained an AUC value of 0.942, and the Transformer, CNN-Transformer models yielded AUC values of 0.957 and 0.962 respectively. Self-attention proved beneficial for learning long-range temporal relationships in gait sequences, speech features, and wearable sensor windows, respectively, in this improvement. The signal changes in the brain may not always be found in isolation and may be associated with abnormalities in Parkinson's disease. Rather, they can emerge over reoccurring steps, prolonged phrases, tremor runs or transitions in movement. Transformer encoders can be used for this kind of pattern as they enable the model to make comparisons over the entire pattern sequence instead of only short patterns locally. This is aligned with the current relevance in the field of sequence modelling for Transformer-based time-series analysis and Parkinson's disease [14], [15], [42], [43].

The best result was achieved with the incorporation of the auxiliary neuroimaging branch, which brought the AUC from 0.965 to 0.973 and accuracy from 93.1% to 94.3%. The improvement was small but

clinically significant as it was an evidence type that could not be measured by gait, speech, or wearable signals. Digital biomarkers are measures that describe a change in behavior and function of people with Parkinson's disease and other MRI and dopamine transporter imaging can provide biological or anatomical information about the neurodegeneration and dopamin deficit. The imaging-only auxiliary model performed well as well with an AUC of 0.934, indicating that the neuroimaging-based features provide diagnostic information. This is consistent with previous research indicating that imaging and clinical data are useful tools to diagnose and early detect Parkinson's disease [17]–[41]. Simultaneously, the results indicate that imaging is not to be considered as a substitute for digital biomarkers, but rather as a complementary validation branch to the primary non-image multimodal model, which was able to reclassify with very good performance.

The multimodal design is substantiated by the ablation analysis. Worn sensor data removal resulted in the greatest decrease in AUC and removing gait and speech resulted in the smallest decrease in AUC. This shows that signals from the wearable sensors were essential to this first model, presumably because they were able to reflect information on dynamic movement which was closely related to the motor impairment. Imaging was added to the secondary model, and there was a slight increase in the AUC from 0.973 to 0.965, indicating that neuroimaging added some value to the model, but did not dominate the prediction. It's a balance that is crucial to the positioning of the work. Medical imaging improves the validation and interpretability for the audience of a medical imaging journal, but the study continues to focus on gait, speech, and wearable digital biomarkers.

The attention-based interpretability results also help to demonstrate the clinical usefulness of the model. For the primary non-image model, the data from the wearables was 38%, gait was 34%, and speech was 28%. The distribution of contribution in the imaging-supported model was subsequently altered to 29% of the contribution from the imaging, 27% from a wearable sensor, 24% from gait, and 20% from speech. The model was not based on a single, predominant modality as evidenced by these values. It rather spread its efforts over imaging and digital biomarkers, which is good for a heterogenous disease like Parkinson's. The clinician reading such an output would be able to tell if the prediction was based primarily on movement impairment, speech change, or the neuroimaging. This is especially crucial because a good model that is not interpretable may not be easily believed in a clinical context.

Both the multimodal attention-based Transformer and the best single-modality wearable model were compared using the statistical comparison, with the multimodal attention-based Transformer achieving a better performance. The imaging-supported multimodal Transformer also exhibited a statistically significant improvement over the primary multimodal Transformer, although the gain was smaller. The results indicate that the proposed attention-based fusion approach is not just an increase in data, but a learning of useful relationships between the modalities. This is further supported by the improvement achieved with feature level and decision level fusion. Feature concatenation and probability averaging are good starting points, but don't capture the varying degree of importance of each modality across participants. Attention-based fusion overcomes this issue by allowing weights of each modality to change based on the input pattern.

The critical point from a clinical point of view is that the best model has a low false-negative rate. For Parkinson's disease detection, false negative results are more worrying than false positive results as they may result in a delay in neurological assessment and monitoring. The most successful imaging-based model misclassified 11 PD samples as healthy control, and 9 healthy control samples as PD. While no AI model is meant to be used as a standalone diagnostic tool, this error pattern indicates that the framework could be helpful as a decision support for early screening, referral support or for long-term monitoring. It should be used to support clinical diagnosis, rather than as a substitute.

The multi-dataset design not only enhances the translatability of the study but also presents a challenge in interpretation. The datasets were gathered from various cohorts and protocols, so the fusion results are intended to be interpreted as representation-level and decision-level multimodal fusion, and not as a full match participant-level multimodal fusion. This design is also methodologically sound and practical, as the datasets of Parkinson's disease are not always available for public use, and they lack joint records of gait, speech, wearable, and imaging data for the same subjects. Multiple public datasets further enhance the reproducibility and enable testing of modality-specific generalization, though future research should confirm the model with matched multimodal datasets with a cohort of gait, speech, wearable and imaging data.

The results validate the importance of a multimodal AI approach supported by medical imaging in the detection of Parkinson's disease. The main contribution is the fusion of gait, speech and wearable sensor data by embedding them into an attention-based Transformer model with the neuroimaging branch providing secondary medical imaging relevance. This is a structure which retains the original digital biomarker objective and continues to offer a meaningful imaging supported validation pathway. Such a

framework could benefit future clinical decision support systems where functional signals, behavioral features and imaging data are all integrated for more accurate early detection of Parkinson's disease. The section that follows addresses various limitations on the datasets, including dataset heterogeneity, modality mismatch, imaging availability, external validation and prospective testing in matched multimodal datasets.

## 6. Limitations

The results should be considered with certain restraints. The first limitation is the use of a multi-dataset design. Gait, speech, and wearable sensor data were used from publicly available datasets, as were tasks performed with the smartphone and auxiliary neuroimaging support. This enhances accessibility and reproducibility, although the data were collected in different cohorts, devices, protocols and clinical contexts. So the findings must be interpreted as a “harmonised multi-dataset analysis” and not as a “true” participant-level multi-modal study. Multimodal direct fusion would be more convincing if obtained from a single cohort, consisting of gait, speech, wearable recordings, and neuroimaging data from the same participants.

Another significant drawback is the data heterogeneity. Gait recordings, speech signals, wearable sensor data, task data on smartphones and MRI or DaTscan images vary with respect to sampling rate, length of recording, signal quality, location of the device, and preprocessing requirements. Normalization, segmentation and feature harmonization were performed but there is a possibility of some variation in the set which may affect the model learning. This is particularly important for data collected via wearable and smartphone devices, where the variability of movement in the real world, the participant's compliance to wearing the device, and the position of the sensor can impact the quality of the data.

There are also limitations to the practical applications of the auxiliary imaging branch. While medical imaging was not the primary modality of the original framework, PPMI neuroimaging data was added to enhance medical imaging relevance and for secondary validation. While both are useful anatomical and biological data, MRI is more expensive, less common, and less repeatable than gait or speech or data from wearable sensors. Thus, the model presented with images should be considered as complementary to the main digital biomarker rather than an alternative to it.

Another limitation of the proposed architecture is that it fails to implement all of the operations. Transformer models can model long-range dependencies, but typically need more data and more computational power than CNN and LSTM models. With smaller or imbalanced data sets, there is the possibility of overfitting. This risk was mitigated by regularization, early stopping, class weighting and a cross-validation, but external validation to independent prospective cohorts is still required.

Generalizability may also be impacted by class imbalance and ethnic and racial diversity. Datasets of Parkinson's disease cases and healthy controls are available in the public domain but may not be equal in terms of the number of cases and controls, age, sex, disease duration, medication status, severity stage and geographic background. Balanced accuracy, sensitivity and specificity along with accuracy were used, and future research could have a breakdown of outcomes by age, disease stage, medication state, and demographic.

The outputs of the explainability should be used with caution. Transparency scores such as attention scores, saliency maps, Grad-CAM outputs and modality contribution scores help increase transparency, but do not indicate clinical causality. A high attention weight means that the model used more the modality or region when predicting the data, but does not necessarily mean that the highlighted feature is the biological cause of Parkinson's. These outputs should aid in clinical interpretation and should be interpreted with expert judgment.

The framework needs to be validated in the future in a prospective study with matched multimodal data, recorded according to standard protocols, to achieve clinical translation. Robustness should be tested in future studies with other hospitals, imaging scanners, wearable devices, speech recording environments and in other patient groups. To be suitable for clinical decision support, real-world screening, or longitudinal monitoring of Parkinson's disease, such validation is needed prior to use in the real world.

## 7. Potential Clinical and Research Implications and Future Directions

The proposed framework could have potential clinical relevance as it integrates gait, speech, wearable sensor and auxiliary neuroimaging data, as opposed to relying on individual sources of evidence. Patients with Parkinson's disease are heterogeneous and a single biomarker cannot fully represent the pattern of the disease. Gait features can be used as indicators of locomotor difficulty, speech features can be used as indicators of vocal and/or articulatory changes, wearable sensors can be used to record tremor and movement variability, and neuroimaging can be used as supporting evidence of structural and/or

dopaminergic changes. The framework can incorporate these signals to achieve earlier detection of those who need a detailed neurological assessment.

The model may also be used as a future screening and monitoring tool as opposed to a replacement for neurologists from a clinical decision support standpoint. It can reveal the relative contributions of gait, speech, wearable sensors, and imaging-derived features to a prediction with its outputs being easy to interpret. The high Parkinson's disease prediction, for instance, with the help of wearable tremor patterns and gait irregularity is more clinically relevant than the black-box prediction.

Improving the relevance of the framework for medical imaging-supported clinical workflows by the auxiliary neuroimaging branch. Biological and anatomical context can be provided by MRI or DaTscan/SPECT-derived features, and real-world function is provided by digital biomarkers. This may be helpful in cases of poor clinical features, early disease, and when distinguishing from other types of movement disorders. The imaging branch is not intended to supplant the core digital biomarker package, but can be used as a supplementary validation for secondary use and improve interpretation of the models.

The framework also has implications for remote monitoring, digital health research. Tasks on the smartphone and wearable sensors can be repeated over time to see changes in movement and speech OUT of the clinic. This can help with monitoring disease progression and medication response, as well as motor fluctuation over time. The non-image multimodal model can also work, if the neuroimaging is not easily available, with gait, speech and wearable sensors. When MRI imaging or imaging of the dopamine transporters is available, specialized analysis can be added in specialized centers.

This work suggests the necessity of having standardized and matched multimodal Parkinson's disease datasets for future research. Existing public datasets are helpful but often use other cohorts and/or other protocols. Future studies should gather gait, speech, wearable signals, clinical scores, and neuro-imaging data within the same participants to enable further fusion of participants at the individual level and more reliable validation across disease stages and demographic groups.

The suggested strategy also enables the development of future explainable and clinically viable AI systems. The attention scores, modality contribution values, the temporal saliency maps and imaging heat maps can provide information on which ones are the most informative biomarkers. These tools can help identify if there are differential diagnostic signals, like subtle imaging changes in early disease vs changes in speech in early disease or gait and wearable sensor changes in advanced disease.

Prospective validation, external testing and integration in clinical workflow in other hospitals, imaging scanners, wearable devices, speech recording environments and patient groups should now be the next focus. Other factors like data collection time, patient comfort, computational cost, privacy and ease of use for the clinician should also be assessed. If validated carefully, the medical imaging assisted multimodal AI system could serve as a potential basis for additional studies of screening, monitoring, and decision making for Parkinson's disease.

## 8. Conclusion

This study proposed a multimodal artificial intelligence (AI) system that leverages wearable sensor data and speech and gait data as primary digital biomarkers with Parkinson's disease (PD) validation branch based on neuroimaging from the Parkinson's Progression Markers Initiative (PPMI). Combining functional, behavioral, and imaging data eliminated the weaknesses of relying on a single modality AI model and led to a more complete characterization of the changes associated with PD. The multimodal attention-based Transformer model outperformed individual gait-only, speech-only, and wearable-only model, and the imaging-supported model had the highest performance, indicating that the imaging modality could be used as a support modality to the primary digital biomarker framework rather than a replacement. The clinical relevance was provided by the explainability layer which demonstrated the importance of each modality by means of attention scores, saliency patterns, and by explaining the interpretation in an imaging-based way. The proposed framework should, however, be considered as a decision support approach, and not as a substitute for a neurological diagnosis. Prospective matched, multimodal data sets are needed to validate the approach in the future, for use in Parkinson's disease screening, follow-up, and medical imaging-aided clinical decision support.

## REFERENCES

[1] Y. Ben-Shlomo, S. Darweesh, J. Llibre-Guerra, C. Marras, M. San Luciano, and C. Tanner, “The epidemiology of Parkinson’s disease,” The Lancet, vol. 403, no. 10423, pp. 283–292, Jan. 2024, doi: 10.1016/S0140-6736(23)01419-8.

[2] D. Su et al., "Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021," BMJ, vol. 388, p. e080952, Mar. 2025, doi: 10.1136/bmj-2024-080952.
[3] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," The Lancet, vol. 397, no. 10291, pp. 2284–2303, Jun. 2021, doi: 10.1016/S0140-6736(21)00218-X.
[4] C. E. Yau, E. C. K. Ho, N. Y. Ong, C. J. K. Loh, A. S. Mai, and E. Tan, "Innovative technology-based interventions in Parkinson's disease: A systematic review and meta-analysis," Ann Clin Transl Neurol, vol. 11, no. 10, pp. 2548–2562, Oct. 2024, doi: 10.1002/acn3.52160.
[5] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, "Challenges in the diagnosis of Parkinson's disease," The Lancet Neurology, vol. 20, no. 5, pp. 385–397, May 2021, doi: 10.1016/S1474-4422(21)00030-2.
[6] T. Simuni et al., "A biological definition of neuronal α-synuclein disease: towards an integrated staging system for research," The Lancet Neurology, vol. 23, no. 2, pp. 178–190, Feb. 2024, doi: 10.1016/S1474-4422(23)00405-2.
[7] A.-K. Schalkamp, K. J. Peall, N. A. Harrison, and C. Sandor, "Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis," Nat Med, vol. 29, no. 8, pp. 2048–2056, Aug. 2023, doi: 10.1038/s41591-023-02440-2.
[8] L. Sigcha et al., "Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review," Expert Systems with Applications, vol. 229, p. 120541, Nov. 2023, doi: 10.1016/j.eswa.2023.120541.
[9] A. Suppa et al., "Voice in Parkinson's Disease: A Machine Learning Study," Front. Neurol., vol. 13, p. 831428, Feb. 2022, doi: 10.3389/fneur.2022.831428.
[10] L. Borzì, L. Sigcha, D. Rodríguez-Martín, and G. Olmo, "Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor," Artificial Intelligence in Medicine, vol. 135, p. 102459, Jan. 2023, doi: 10.1016/j.artmed.2022.102459.
[11] M. B. Makarious et al., "Multi-modality machine learning predicting Parkinson's disease," npj Parkinsons Dis., vol. 8, no. 1, p. 35, Apr. 2022, doi: 10.1038/s41531-022-00288-w.
[12] V. Dentamaro, D. Impedovo, L. Musti, G. Pirlo, and P. Taurisano, "Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database," Sci Rep, vol. 14, no. 1, p. 20941, Sep. 2024, doi: 10.1038/s41598-024-70165-4.
[13] A. Zhao, Y. Liu, X. Yu, X. Xing, and H. Zhou, "Artificial Intelligence-Enabled Detection and Assessment of Parkinson's Disease Using Multimodal Data: A Survey," 2025, SSRN. doi: 10.2139/ssrn.5167417.
[14] L. Sigcha et al., "Improvement of Performance in Freezing of Gait detection in Parkinson's Disease using Transformer networks and a single waist-worn triaxial accelerometer," Engineering Applications of Artificial Intelligence, vol. 116, p. 105482, Nov. 2022, doi: 10.1016/j.engappai.2022.105482.
[15] S. Naimi, W. Bouachir, and G.-A. Bilodeau, "1D-Convolutional transformer for Parkinson disease diagnosis from gait," Nov. 06, 2023, arXiv: arXiv:2311.03177. doi: 10.48550/arXiv.2311.03177.
[16] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," Nat Med, vol. 28, no. 9, pp. 1773–1784, Sep. 2022, doi: 10.1038/s41591-022-01981-2.
[17] K. Marek et al., "The Parkinson Progression Marker Initiative (PPMI)," Progress in Neurobiology, vol. 95, no. 4, pp. 629–635, Dec. 2011, doi: 10.1016/j.pneurobio.2011.09.005.
[18] S. R. Suwijn, C. J. Van Boheemen, R. J. De Haan, G. Tissingh, J. Booij, and R. M. De Bie, "The diagnostic accuracy of dopamine transporter SPECT imaging to detect nigrostriatal cell loss in patients with Parkinson's disease or clinically uncertain parkinsonism: a systematic review," EJNMMI Res, vol. 5, no. 1, p. 12, Dec. 2015, doi: 10.1186/s13550-015-0087-1.
[19] J. Zhang, "Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease," npj Parkinsons Dis., vol. 8, no. 1, p. 13, Jan. 2022, doi: 10.1038/s41531-021-00266-8.
[20] T. Welton et al., "Classification of Parkinson's disease by deep learning on midbrain MRI," Front. Aging Neurosci., vol. 16, p. 1425095, Aug. 2024, doi: 10.3389/fnagi.2024.1425095.
[21] A. Mirelman et al., "Gait impairments in Parkinson's disease," The Lancet Neurology, vol. 18, no. 7, pp. 697–708, Jul. 2019, doi: 10.1016/S1474-4422(19)30044-4.
[22] J. M. Hausdorff, "Gait dynamics in Parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 19, no. 2, p. 026113, Jun. 2009, doi: 10.1063/1.3147408.
[23] V. V. Shah et al., "Digital Biomarkers of Mobility in Parkinson's Disease During Daily Living," JPD, vol. 10, no. 3, pp. 1099–1111, Jul. 2020, doi: 10.3233/JPD-201914.
[24] M. Mancini et al., "Digital gait biomarkers in Parkinson's disease: susceptibility/risk, progression, response to exercise, and prognosis," npj Parkinsons Dis., vol. 11, no. 1, p. 51, Mar. 2025, doi: 10.1038/s41531-025-00897-1.

[25] B. M. Bot et al., "The mPower study, Parkinson disease mobile data collected using ResearchKit," Sci Data, vol. 3, no. 1, p. 160011, Mar. 2016, doi: 10.1038/sdata.2016.11.
[26] L. Lonini et al., "Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models," npj Digital Med, vol. 1, no. 1, p. 64, Nov. 2018, doi: 10.1038/s41746-018-0071-z.
[27] J. L. Adams et al., "Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study," npj Parkinsons Dis., vol. 9, no. 1, p. 64, Apr. 2023, doi: 10.1038/s41531-023-00497-x.
[28] F. Lipsmeier et al., "Reliability and validity of the Roche PD Mobile Application for remote monitoring of early Parkinson's disease," Sci Rep, vol. 12, no. 1, p. 12081, Jul. 2022, doi: 10.1038/s41598-022-15874-4.
[29] S. K. Sieberts et al., "Developing better digital health measures of Parkinson's disease using free living data and a crowdsourced data analysis challenge," PLOS Digit Health, vol. 2, no. 3, p. e0000208, Mar. 2023, doi: 10.1371/journal.pdig.0000208.
[30 R. Powers et al., "Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease," Sci. Transl. Med., vol. 13, no. 579, p. eabd7865, Feb. 2021, doi: 10.1126/scitranslmed.abd7865.
[31] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," IEEE Trans. Biomed. Eng., vol. 59, no. 5, pp. 1264–1271, May 2012, doi: 10.1109/TBME.2012.2183367.
[32] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1015–1022, Apr. 2009, doi: 10.1109/TBME.2008.2005954.
[33] C. O. Sakar et al., "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," Applied Soft Computing, vol. 74, pp. 255–263, Jan. 2019, doi: 10.1016/j.asoc.2018.10.022.
[34] C. Quan, K. Ren, Z. Luo, Z. Chen, and Y. Ling, "End-to-end deep learning approach for Parkinson's disease detection from speech signals," Biocybernetics and Biomedical Engineering, vol. 42, no. 2, pp. 556–574, Apr. 2022, doi: 10.1016/j.bbe.2022.04.002.
[35] L. Van Gelderen and C. Tejedor-García, "Innovative Speech-Based Deep Learning Approaches for Parkinson's Disease Classification: A Systematic Review," Applied Sciences, vol. 14, no. 17, p. 7873, Sep. 2024, doi: 10.3390/app14177873.
[36] M. A. Hossain and F. Amenta, "Machine Learning-Based Classification of Parkinson's Disease Patients Using Speech Biomarkers," Journal of Parkinson's Disease, vol. 14, no. 1, pp. 95–109, Jan. 2024, doi: 10.3233/JPD-230002.
[37] M. Junaid, S. Ali, F. Eid, S. El-Sappagh, and T. Abuhmed, "Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 234, p. 107495, Jun. 2023, doi: 10.1016/j.cmpb.2023.107495.
[38] Godala, Sravanthi, and M. Sunil Kumar. "A weight optimized deep learning model for cluster based intrusion detection system." Optical and Quantum Electronics 55.14 (2023): 1224.
[39] Suman, J. V., Mahammad, F. S., Sunil Kumar, M., Sai Chandana, B., & Majji, S. (2024). Leveraging natural language processing in conversational AI agents to improve healthcare security. Conversational Artificial Intelligence, 699-711.
[40] Gandikota, Hari Prasad, Abirami S, and Sunil Kumar M. "CT scan pancreatic cancer segmentation and classification using deep learning and the tunicate swarm algorithm." Plos one 18.11 (2023): e0292785.
[41] Kumar, M. S., & Harshitha, D. (2019). Process innovation methods on business process reengineering. Int. J. Innov. Technol. Explor. Eng, 8(11), 2766-2768.Wolpert, D. H. (2023).
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 5998-6008.
[43] Q. Wen et al., "Transformers in Time Series: A Survey," in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 6778–6786. doi: 10.24963/ijcai.2023/759.