Preprints.org

Article

A Comparative Analysis of Parkinson's Disease Diagnosis Approaches Using Drawing-Based Datasets: Utilizing Large Language Models, Machine Learning, and Fuzzy Ontologies

Adam Koletis [*] , Pavlos Bitilis , Georgios Bouchouras , Konstantinos Kotis [*]

Article

A Comparative Analysis of Parkinson's Disease Diagnosis Approaches Using Drawing-Based Datasets: Utilizing Large Language Models, Machine Learning, and Fuzzy Ontologies

Adam Koletis [1],*, Pavlos Bitilis [2], Georgios Bouchouras [3] and Konstantinos Kotis [4],*

[1] Intelligent Systems Lab, Department of Cultural Technology and Communication, University of the Aegean, 81100 Mytilene, Greece

[2] University of the Aegean

[3] University of the Aegean

[4] University of the Aegean

* Correspondence: ctd21001@ct.aegean.gr (A.K.); kotis@aegean.gr (K.K.)

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder that impairs motor function, often causing tremors and difficulty with movement control. A promising diagnostic method involves analyzing hand-drawn patterns, such as spirals and waves, which show characteristic distortions in individuals with PD. This study compares three computational approaches for classifying individuals as Parkinsonian or healthy based on drawing-derived features: (1) Large Language Models (LLMs), (2) traditional machine learning (ML) algorithms, and (3) a fuzzy ontology-based method using fuzzy sets and Fuzzy-OWL2. Each method offers unique strengths: LLMs leverage pre-trained knowledge for subtle pattern detection, ML algorithms excel in feature extraction and predictive accuracy, and fuzzy ontologies provide interpretable, logic-based reasoning under uncertainty. Using three structured handwriting datasets of varying complexity, we assessed performance in terms of accuracy, interpretability, and generalization. Among the approaches, the fuzzy ontology-based method showed the strongest performance on complex tasks, achieving a high F1-score, while ML models demonstrated strong generalization and LLMs offered a reliable, interpretable baseline. These findings suggest that combining symbolic and statistical AI may improve drawing-based PD diagnosis.

Keywords: Parkinson's Disease (PD); hand-drawn pattern analysis; Machine Learning (ML); Large Language Models (LLMs); fuzzy ontology

---

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily affects movement. It is characterized by the degeneration of dopaminergic neurons in the substantia nigra region of the brain, leading to motor symptoms such as tremors, bradykinesia (slowness of movement), rigidity, and postural instability. In addition to motor impairments, many patients also experience non-motor symptoms including cognitive decline, sleep disturbances, and mood disorders. The exact cause of Parkinson's remains unclear, though genetic and environmental factors are believed to contribute to its development [1]. PD reveals paradoxical effects on creativity. While motor impairments and cognitive decline are well-known features, a subset of patients demonstrates either newly emergent or significantly increased artistic behavior, especially in the visual arts. In some cases, artistic production begins for the first time in late adulthood, often after the initiation of dopaminergic therapy [2,3]. This "creative drive" is not simply compulsive behavior but a structured

and meaningful expression of visual creativity, often described as fulfilling and pleasurable [2,4]. The phenomenon is frequently linked to dopamine agonists, which stimulate the mesolimbic pathways involved in motivation and reward [2,3].

Neurologically, increased artistic behavior is believed to result from disinhibition of the right hemisphere, particularly its visuospatial processing areas, due to dopamine stimulation or unilateral functional imbalance [4,5]. This aligns with the concept of paradoxical functional facilitation (PFF) where injury or degeneration in one area (e.g., frontal lobes or left hemisphere) may "release" latent capacities in another (e.g., visuospatial right hemisphere) [6,7]. Functional imaging studies have supported this model. [3,5] described a case where a patient transitioned from abstract to hyper realistic painting even before motor symptoms appeared, suggesting early visual-cognitive alterations.

Artistic themes in PD patients tend to shift toward vivid, detailed, and often highly structured forms, even as motor skills decline. Remarkably, tremors and bradykinesia often disappear during painting sessions [3,8–12]. fMRI studies have shown that skilled artists, compared to novices, activate higher-order frontal areas, suggesting that PD patients who develop or maintain creativity may rely on compensatory executive or associative pathways [13].

Historical and iconographic examples suggest that Parkinsonian features may have been represented artistically long before the clinical definition of the disease. These depictions include self-portraits by Albrecht Dürer [14] and expressive dystonic imagery in the works of Pieter Bruegel [15]. Such visual traditions highlight how neurological conditions have been internalized and reflected through art across centuries.

Frontotemporal dementia (FTD), particularly its semantic and behavioral variants, is also associated with emergent visual creativity. Patients with left anterior temporal lobe atrophy may begin painting for the first time, often producing detailed and emotionally expressive works. This is attributed to the release of right posterior cortical functions due to degeneration in the language-dominant left hemisphere [5–7]. One case study [7] documented an artist whose work became freer and more original despite a progressive aphasia syndrome, reinforcing the notion that language inhibition can unleash visual creativity.

In contrast, Alzheimer's disease (AD) typically leads to a decline in artistic output. During the early stages, some patients retain enough visual and procedural memory to continue creating simple artworks. Over time, however, their creations become increasingly abstract and less detailed, reflecting cognitive and visuospatial deterioration [6,7]. The late works of Willem de Kooning are frequently cited as a poignant example of this artistic evolution, marked by a simplification of complex structures and forms into emotive strokes [6,7,16].

Corticobasal degeneration and other related syndromes can result in significant changes in artistic production. One professional artist with corticobasal degeneration experienced increasing distortion and chaos in their artwork, particularly on the side contralateral to the brain lesion. The resulting features such as disinhibition, neglect, visual-spatial impairment closely mirrored the patient's neuropsychological profile and imaging findings [16].

The intersection of Parkinson's disease and art illuminates an unexpected domain where neurodegeneration may facilitate, rather than impair, certain creative capacities. Through the modulation of dopaminergic systems and lateralized brain function, some patients develop a compelling creative drive. This phenomenon is not exclusive to PD; FTD also reveals emergent visual abilities, while AD and other diseases often impair artistic production. Collectively, these findings suggest that artistic behavior is a sensitive and multifaceted reflection of the evolving brain.

In recent years, the digitization of handwriting and advances in ML (ML) have enabled new computational approaches for analyzing these motor patterns. Supervised learning models and deep neural networks have demonstrated notable predictive power in PD classification tasks [17,18]. However, such approaches often suffer from limited interpretability and lack adaptability in real-world clinical workflows, two factors critical for medical decision-making.

This paper presents a comprehensive comparative study of three distinct approaches to PD detection using hand-drawn pattern analysis:

- Large Language Models (LLMs): ChatGPT is used for classifying PD and healthy individuals based on structured handwriting features.

- Fuzzy Ontology-Based Reasoning: Implemented in Protégé using OWL and fuzzy logic, this method supports human-readable, rule-based decision-making.

- Traditional Supervised ML Models: Including Random Forest classifiers trained on dynamic and kinematic handwriting signal features.

Our study evaluates each approach across three custom-designed datasets with increasing real-world complexity: (1) a clean set of statistical handwriting features, (2) a fuzzy-annotated dataset incorporating expert linguistic rules, and (3) real-time captured spiral and meander drawings collected via digitizing tablets [19–21].

This work is driven by the need to bridge the gap between predictive accuracy and interpretability in clinical AI systems [20,21]. While LLMs like ChatGPT represent a novel class of general-purpose inference tools, their reliability in structured classification settings for clinical data remains underexplored. In contrast, fuzzy ontology systems offer transparent decision-making frameworks but often lack robustness and generalization.   Traditional ML models sit between these extremes, balancing performance and interpretability through optimized data-driven learning. While they often achieve high accuracy on structured datasets, their transparency and adaptability in uncertain or noisy real-world environments may be limited [22].

Our key contributions include a unified evaluation framework for these three reasoning paradigms applied to PD handwriting data; the design of three tiered datasets representing escalating complexity and realism; and a quantitative comparison using standard metrics such as F1-score, accuracy, and confusion matrices.

We hypothesize that:

1. Supervised machine learning (ML) models are expected to achieve the highest classification accuracy, due to their ability to optimally learn from structured and clean feature sets.

2. The fuzzy ontology-based system is expected to provide the greatest interpretability and robustness, particularly in scenarios involving uncertainty or incomplete information, thanks to its use of expert-defined symbolic rules.

3. The large language model (LLM) is anticipated to serve as a competitive baseline, effectively bridging human-readable reasoning with numeric feature analysis, and offering moderate performance with high usability.

By testing these hypotheses across datasets of increasing complexity, this study aims to evaluate the accuracy, explainability, and adaptability in PD diagnosis systems

In summary, this paper presents a comparative evaluation of three classification paradigms, traditional ML, fuzzy ontologies, and LLMs, for detecting Parkinson's disease from handwriting data. After introducing relevant background (Section 2), we describe the related work (Section 3), followed by the methodology and model implementation (Section 4). Section 5 presents experimental results, and Section 6 is dedicated to discussion and limitations, while Section 7 concludes with insights and future directions.

2. Background Knowledge

2.1. Parkinson's Disease and Artistic Behavior

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder characterized by a combination of motor and non-motor symptoms. Motor symptoms include tremor, rigidity, bradykinesia (slowness of movement), and postural instability, largely due to the degeneration of dopaminergic neurons in the substantia nigra [23]. Non-motor symptoms may include cognitive impairment, mood disturbances, and sleep dysfunction, complicating diagnosis and management.

Interestingly, a subset of PD patients exhibits increased or newly emergent artistic expression, particularly in the visual arts. This paradoxical phenomenon has been linked to dopaminergic therapy and right-hemisphere disinhibition, which may enhance visuospatial and creative capacities [24]. Art-based tasks such as spiral drawing or wave tracing have therefore gained diagnostic interest, as they capture nuanced aspects of spatial control and fine motor coordination.

2.2. Feature Extraction from Handwriting and Drawing Tasks

Handwriting and sketch-based tasks are widely used in PD research due to their ability to reveal subtle motor abnormalities. From these signals, several features are extracted to quantify movement quality and structure [25]. Key metrics include:

- Entropy: Measures complexity or randomness of motion.
- Smoothness: Reflects the fluidity of movement; lower values suggest tremor or rigidity.
- Kurtosis: Captures the sharpness or flatness of motion peaks.
- Wave Energy: Represents the total signal energy during a stroke.
- Average Pressure: Measures the pen tip's applied force.
- Average Speed: Reflects overall drawing velocity.
- Segment Length: Tracks the length of individual stroke segments.

These features serve as input to machine learning models and fuzzy logic systems for PD classification.

2.3. Supervised Machine Learning

Supervised machine learning (ML) is a fundamental approach in PD diagnosis, using labeled datasets to train predictive models that classify individuals based on extracted features [26]. Popular classifiers include support vector machines, decision trees, and ensemble methods such as Random Forest.

Random Forest, introduced by Breiman [27], is an ensemble method that builds multiple decision trees and averages their outputs to improve generalization and reduce overfitting. It is well-suited to biomedical applications due to its ability to handle high-dimensional and non-linear data. In this study, the Random Forest implementation from the scikit-learn library was used across all datasets [28].

2.4. Large Language Models (LLMs)

Large Language Models (LLMs), such as GPT-based models like ChatGPT, are transformer-based deep learning architectures trained on vast textual datasets. In healthcare, they have been applied to tasks such as clinical summarization, question answering, patient support, and research assistance [29]. While LLMs offer powerful generalization and human-like interaction, they also have limitations such as hallucination of facts, reduced accuracy in domain-specific tasks, and a lack of native support for structured signal data unless fine-tuned.

In this study, ChatGPT was used to classify PD samples through rule-based logic derived from structured handwriting features, offering a human-readable reasoning baseline.

2.5. Fuzzy Ontologies and Semantic Reasoning in Protégé

Fuzzy ontologies are knowledge representation tools that incorporate vagueness and partial truth values into logical reasoning systems. Unlike classical ontologies that rely on binary logic, fuzzy ontologies model gradations of membership—making them ideal for representing medical data with overlapping symptoms or uncertain thresholds [30].

In this work, fuzzy ontologies were developed using fuzzyOWL2 and deployed in the Protégé environment, with reasoning supported by the HermiT reasoner [31]. Semantic categories such as ParkinsonianSeverityHigh or PressureLow were defined using fuzzy rules and triangular

membership functions. This approach allows for interpretable, rule-based classification that captures uncertainty in patient behavior and motor symptoms.

2.6. Evaluation Metrics and Statistical Testing

To assess both data distribution and model performance, several statistical and evaluation techniques were employed.

The Shapiro–Wilk test was used to evaluate the normality of feature distributions. It is suitable for small to medium-sized datasets and helps determine whether Pearson or Spearman correlation should be used in later analyses.

The Welch's t-test , a robust variation of the independent samples t-test , was applied to compare feature means between the synthetic training data and real-world test data. A p-value below 0.05 indicates statistically significant differences and was used to assess the realism of synthetic data prior to model deployment.

Model performance was evaluated using the F1-score, which balances precision and recall, especially in imbalanced datasets. It is defined as:

$$F_1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

In addition, confusion matrices were used to visualize classification performance by capturing true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) [32].

3. Related Work

3.1. Parkinson's Disease and Artistic Behavior

Research has revealed that some individuals with Parkinson's disease (PD) exhibit changes in artistic abilities, often marked by an increase in creativity or artistic output after the onset of the disease or during dopamine replacement therapy. [33] discusses numerous PD patients who developed newfound artistic tendencies, attributing these changes to dopaminergic stimulation and the modulation of the mesolimbic system, which is associated with reward and creativity. Similarly, [34] report increased artistic production in PD patients undergoing dopamine therapy, supporting the link between neurochemical modulation and visual creativity.

[35] and [36] explore how PD affects visual perception and aesthetic evaluation. Altered sensory processing, such as changes in color perception and spatial awareness, may lead patients to explore new artistic styles or reinterpret traditional motifs. This phenomenon is further documented by [5,37], who describe cases of de novo art creation in patients with advanced PD, suggesting that both the progression of disease and its treatment play integral roles.

[38,39] propose ontology-based systems to model the complex physiological and cognitive states in PD, emphasizing the integration of biomarkers, neuroimaging, and patient behavior for early diagnosis and monitoring. Such models could eventually include creative output as a diagnostic feature or therapeutic marker.

3.2. Art and Other Neurodegenerative Diseases

Alzheimer's disease and frontotemporal dementia (FTD) also show a relationship with altered or emergent artistic expressions. For instance, [40,41] document emergent creativity' in patients with semantic variant primary progressive aphasia, a form of FTD. This phenomenon involves new or intensified creative behavior, which some authors interpret as disinhibition of neural circuits related to executive control.

In Alzheimer's, art therapy is often used to maintain cognitive function and emotional expression, as discussed in 'Art and Alzheimer Type Dementia (various authors). Ontology-driven and fuzzy logic-based systems, such as those by [42] and [43], help classify and interpret behavioral changes, including emotional and creative shifts, providing support in dementia care and diagnosis.

The relationship between Parkinson's disease and artistic behavior highlights how neurodegenerative processes can unlock or modify creative capacities. While much remains to be understood about the neurobiological underpinnings, interdisciplinary studies spanning neurology, psychology, and the arts suggest that both the disease itself and its treatment can influence aesthetic production and appreciation. This insight not only deepens our understanding of the human brain but also opens avenues for novel therapeutic interventions.

3.3. Intelligent Systems and Ontologies in Neurodegenerative Disease Research

Recent contributions from artificial intelligence (AI), fuzzy logic, and ontology-driven models significantly expand the potential for diagnosing and understanding neurodegenerative diseases, including Parkinson's and Alzheimer's. Systems such as AlzFuzzyOnto [42] and the expert system designed by [44] demonstrate how fuzzy-based ontologies and rule-based inference engines can assist in handling the imprecision inherent in clinical data, improving diagnostic support and decision-making for diseases like Parkinson's, Alzheimer's, and Huntington's. Moreover, advanced ontologies such as Wear4PDmove [24] bridge wearable sensor data and patient health records, enabling real-time monitoring, behavior analysis, and alerting mechanisms tailored to Parkinson's disease patients. These AI-driven tools highlight the emerging role of personalized, semantically enriched knowledge systems in supporting both clinical care and research.

Also, recent advances in artificial intelligence (AI), fuzzy logic, and ontology-driven frameworks have introduced new possibilities in modeling and monitoring neurodegenerative and developmental disorders. Vouros et al. 2024 [45] proposed a hybrid approach, **X-HCOME**, which combines Large Language Models (LLMs) like ChatGPT with expert input to engineer ontologies for Parkinson's Disease (PD) monitoring and alerting. This approach enables the semantic integration of wearable sensor data and electronic health records into Personal Health Knowledge Graphs (PHKGs), facilitating real-time reasoning over critical health events such as freezing of gait or medication omission [46]. In parallel, ML-based predictive modeling has demonstrated strong performance in identifying freezing of gait episodes in PD patients under ON and OFF medication states, using physiological and kinematic features to train XGBoost classifiers. Deep learning approaches, such as recurrent neural networks (RNNs), have also been successfully employed to detect temporal anomalies in neuropsychiatric conditions, supporting the adaptability of time-series models for monitoring irregular PD symptom patterns [47]. Complementing these works, a systematic review by Bouchouras and Kotis [48] on autism spectrum disorder (ASD) detection highlights transferable methodologies involving sensor fusion, IoT integration, and multi-modal AI workflows, further reinforcing the generalizability of such intelligent systems to complex neurological conditions. Collectively, these contributions showcase the evolving role of personalized, semantically driven AI systems in clinical neuroinformatics.

4. Materials and Methods

4.1. Datasets

To ensure a robust and comprehensive evaluation, we utilize three handwriting datasets that reflect increasing levels of real-world complexity. Two datasets(Dataset1Wave, Dataset2Spiral), Spiral Drawings and Wave Drawings, are sourced from the publicly available Parkinson's disease collection curated by K. Mader on Kaggle (https://www.kaggle.com/datasets/kmader/parkinsons-drawings). These datasets contain digitized hand-drawn patterns captured using tablets and are derived from studies by [49], who explored handwriting biometrics as a non-invasive marker for Parkinson's disease. The third dataset (Dataset3Spiral) (https://www.kaggle.com/datasets/team-ai/parkinson-disease-spiral-drawings) [50,51] is a feature-based compilation extracted and aggregated from these raw handwriting samples, enriched with statistical and kinematic descriptors. This dataset serves as a clean, structured foundation for both traditional ML pipelines and inference

by language models. Together, these three datasets allow us to test and compare different reasoning paradigms under control and progressively challenging data scenarios.

As previously described, the fuzzy ontologies were trained and evaluated using the same three handwriting-based datasets, each specifically designed to capture fine motor patterns relevant to Parkinson's disease diagnosis. Dataset-1 included spiral drawings, while Dataset-2 contained wave drawings. Each dataset had 36 training samples and 15 testing samples per class, split evenly between Parkinsonian and healthy subjects. These image datasets were preprocessed to extract dynamic features such as velocity, pressure variation, stroke length, and trajectory entropy, which were subsequently fuzzified and encoded into the ontologies. Dataset-3 consisted of raw pen motion signals in text format, collected during digitized drawing tasks. The data files included 15 samples from health controls and 25 from Parkinsonian subjects. Each file contained pen trajectory and stylus metadata delimited in CSV format with the following columns: X, Y, Z, Pressure, GripAngle, Timestamp, and Test ID. For this study, only the **Test ID = 0** was used, corresponding to the **Static Spiral Test**, in which subjects were instructed to trace a predefined spiral pattern. From these signal recordings, kinematic features such as average speed, segment length, and average pressure were computed and mapped to fuzzy categories within the FSPHD ontology. This consistent preprocessing and feature mapping enabled coherent reasoning across all ontologies despite variability in data modalities and formats.

4.2. Data Recording

The data were recorded [49], using a digital tablet (Wacom Intuos) equipped with a pressure-sensitive stylus, allowing the precise capture of both spatial and dynamic features during the drawing task. Participants traced pre-designed spiral and wave templates at a self-paced speed, and the tablet simultaneously recorded high-resolution PNG images along with detailed motion data. Specifically, the tablet captured the (x, y, z) position coordinates of the pen tip, the applied pen pressure, and the timestamp of each point, enabling the extraction of dynamic features such as drawing speed, acceleration, and pressure variation. Participants were seated comfortably in a quiet, well-lit room and were instructed to use their dominant hand for the task without any time constraints imposed. This setup ensured the preservation of subtle motor symptoms, such as tremor amplitude, bradykinesia, and micrographia, while maintaining consistency across all recordings. The structured recording environment and the rich dynamic data captured by the tablet provided a reliable and detailed foundation for subsequent feature analysis and classification tasks.

4.2.1. Data Augmentation

To further enhance the dataset and mitigate the risk of overfitting due to limited real-world samples, synthetic data generation was employed using deep generative modeling. Feature-based real data extracted from the wave images were first processed using Python, with the SDV (Synthetic Data Vault) library. Metadata for the real dataset (training folders) were automatically detected to specify the data types of each feature. A Conditional Tabular GAN (CTGAN) synthesizer was then trained on the real dataset, with training parameters set to 12,000 epochs and custom learning rates for both generator and discriminator networks. To mitigate the risk of data leakage, all synthetic data was generated exclusively from the training portion of the real dataset. The test set remained untouched throughout synthetic data modeling and evaluation. Additionally, descriptive statistics and visual analyses, including boxplots and group-wise summaries, were used to verify that the synthetic features aligned closely with those of real samples. However, despite these precautions, synthetic data can still introduce artifacts, or overly smooth patterns do not present in genuine motor signals. Therefore, we acknowledge the potential for bias and distribution drift, especially given the relatively small size of the real-world dataset. This risk is addressed to the limitations of our findings. After training, a total of 200,000 synthetic samples were generated, from which 1,000 high-quality synthetic samples per class (Healthy and Parkinsonian) were selected. To validate the quality of the synthetic data, a Random Forest classifier was trained on the real dataset and evaluated on the

synthetic samples. Performance metrics, including a classification report and confusion matrix were generated to assess how closely the synthetic data mimicked the true class distributions. Only the true positive synthetic samples, those correctly classified by the real-data-trained model, were retained for further use. This strategy ensured that only the most realistic synthetic instances were incorporated into subsequent modeling stages, improving the robustness and generalizability of the classification system.

4.2.2. Data Validation

After validating the quality of the synthetic samples, a supervised classification experiment was conducted to differentiate between healthy and Parkinsonian subjects. The true positive synthetic dataset was used, with the target variable encoded as binary labels (0 for Healthy, 1 for Parkinsonian). All relevant features, excluding identifiers such as 'Image' and 'Label', were selected as predictors. The dataset was randomly split into a training set (70%) and a testing set (30%) using stratified sampling to preserve class balance. A Random Forest classifier was trained on the training set and evaluated on the testing set. Performance metrics, including a classification report and confusion matrix, were generated to assess the model's accuracy and class-specific performance. Additionally, feature importance scores were computed to rank the contribution of each feature to the classification decision. To further explore the statistical properties of selected key features (Smoothness, Kurtosis, Signal Entropy, Wave Energy, Irregularities, and Skewness), Shapiro-Wilk tests were performed to assess normality (Figure 1). Depending on the normality result, either Pearson or Spearman correlation coefficients were calculated to examine the relationship between each feature and the disease label. This comprehensive evaluation provided insights into both the predictive value and the statistical relevance of the extracted features for distinguishing between healthy and Parkinsonian subjects.

```
Smoothness          : p = 0.5435 → Same Distribution ✅
Average Amplitude   : p = 0.1767 → Same Distribution ✅
Irregularities      : p = 0.7873 → Same Distribution ✅
Peak Count          : Skipped (not enough variation)
Trough Count        : Skipped (not enough variation)
Zero Crossings      : Skipped (not enough variation)
Wave Energy         : p = 0.4358 → Same Distribution ✅
Signal Entropy      : p = 0.5746 → Same Distribution ✅
Skewness            : p = 0.3554 → Same Distribution ✅
Kurtosis            : p = 0.4395 → Same Distribution ✅
Dominant Frequency  : Skipped (not enough variation)
```

Figure 1. Statistical comparison of synthetic training and real-world test samples in Dataset-1 using the Shapiro–Wilk and Welch's t-test. Features with p-values > 0.05 are considered to follow the same distribution, supporting the statistical realism of the synthetic data. Features with insufficient variation were skipped.
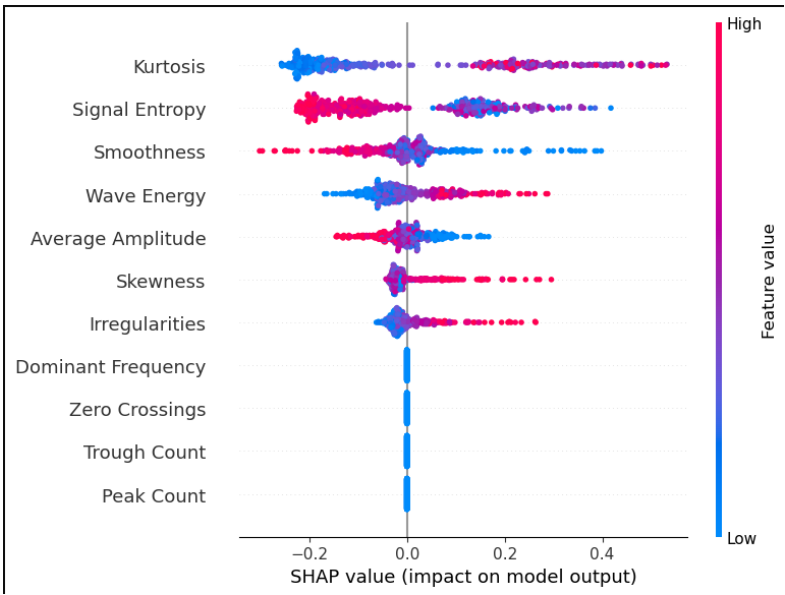
Figure 2. SHAP (SHapley Additive exPlanations) summary plot illustrating the impact of each handwriting feature on the Random Forest model's output for Dataset1. Higher SHAP values indicate greater influence on the model's classification decision, with feature values color-coded from low (blue) to high (red).

4.3. ChatGPT

To assess the potential of large language models (LLMs) in biomedical signal classification, we employed ChatGPT to perform binary classification of Parkinson's disease (PD) versus healthy subjects, using structured numerical features derived from digitized handwriting data. Specifically, we extracted tabular features from spiral and wave drawings, as well as preprocessed statistical datasets. These included signal-based and geometric descriptors such as drawing smoothness, signal entropy, wave energy, average speed, pen pressure variability, and path length.

Each data sample was transformed into a standardized natural language prompt, embedding the numerical values in a consistent clinical format. For example: *"A patient's wave drawing yields an entropy of 0.65, energy of 1.2, and smoothness index of 0.43. Based on these features, does the drawing belong to a Parkinson's patient or a healthy individual?"* ChatGPT was queried in a zero-shot setting—without domain-specific fine-tuning—and its response was parsed into a binary class label ("Parkinson" or "Healthy").

To ensure consistency and validity, we applied this classification procedure to the same held-out real test sets used in the ML and ontology-based approaches. Classification performance was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. Unlike fuzzy ontology models, which use explicit and traceable rules, the ChatGPT-based approach relies on implicit reasoning. This limits interpretability and makes it difficult to analyze errors or identify decision boundaries. Nonetheless, the method offered a lightweight and human-readable approach to infer diagnostic labels from structured data, bridging numerical representation and natural language reasoning.

Recent studies have explored the potential of large language models (LLMs), particularly ChatGPT, in supporting medical and radiologic decision-making, as well as healthcare education. Rao et al. (2023) evaluated ChatGPT as an adjunct in radiologic decision-making and found it showed promise in assisting clinical workflows, albeit with limitations in accuracy and context understanding [52]. Similarly, Singhal et al. (2023) demonstrated that LLMs can encode substantial clinical knowledge, indicating their usefulness in various healthcare applications [53]. In the context of medical education, Kung et al. (2023) assessed ChatGPT's performance on the USMLE and concluded that it could serve as a valuable educational tool for medical students [54]. Complementing these findings, Sallam (2023) conducted a systematic review highlighting ChatGPT's utility across

healthcare education, research, and practice, while also addressing concerns related to ethical use and reliability [29,54].
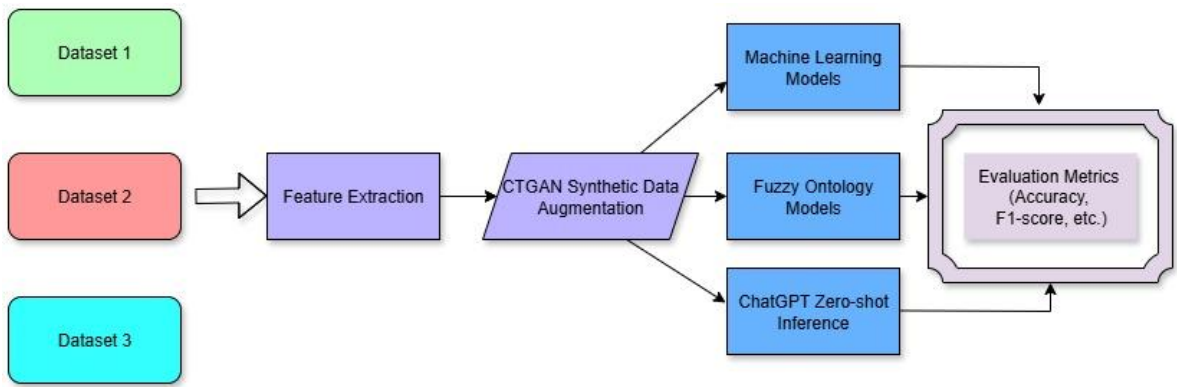
Figure 3. Schematic representation of the experimental pipeline. Handwriting datasets (spiral, wave, and statistical) were processed to extract signal and geometric features. CTGAN was applied for data augmentation. Three classification approaches—traditional ML, fuzzy ontology-based reasoning, and large language models (ChatGPT)—were independently applied to evaluate diagnostic performance using the same test set.

4.4. Fuzzy Ontology

Three fuzzy ontologies, namely, FWP, FSP, and FSPHD (https://github.com/adklts/ParkinsonPredictionMethodsHanddrawn) were developed following a unified design philosophy to ensure interoperability, consistency in interpretation, and methodological comparability. At the core of each ontology lies a fuzzy modeling approach that encodes numerical signal features using **triangular membership functions**, chosen for their simplicity, interpretability, and suitability in capturing gradual transitions typical of human motor behavior. The fuzzy sets for each feature were generated programmatically using a percentile-based function designed to create **tight fuzzy partitions** with partially overlapping regions. Specifically, low, medium, and high fuzzy labels were constructed using percentiles spanning from the 20th to the 80th percentile of each feature's distribution, thereby enabling meaningful semantic fuzzification across diverse data scales. In terms of reasoning, all ontologies employed the **HermiT reasoner** integrated within the Protégé platform. HermiT was selected for its compliance with OWL 2 DL standards, robust support for consistency checking, and proven compatibility with fuzzy ontology plugins such as **FuzzyOWL2**, which allows the interpretation of fuzzy datatypes, modifiers, and rule-based inference over imprecise knowledge. Prior studies have demonstrated HermiT's efficiency and correctness in complex semantic models involving fuzzy extensions and clinical data [55]. Finally, the ontologies maintain a **common structural backbone**, including the top-sublevel classes Parkinsonian and Healthy under the top-level class Individuals, each containing subclasses to represent condition severity, namely ParkinsonianSeverityHigh, Medium, and Low, as well as HealthyStatusHigh, Medium, and Low. This uniform class schema ensures alignment in the interpretability of inference results across different biometric domains and facilitates comparative evaluations of the underlying fuzzy logic rules.

4.4.1. FWP Wave

The FWP ontology was developed to model and reason over fuzzy signal characteristics extracted from wave drawing tasks performed by Parkinsonian and healthy individuals. The ontology defines a structured class hierarchy under the root class Thing, the top-level class Individuals and central class Measurement captures feature-level signal representations through data properties. These include hasKurtosis, hasSignalEntropy, and hasSmoothness, all typed with the double datatype and linked to individuals representing drawing measurements. Fuzzy datatypes were defined for each signal dimension using triangular membership functions within the

FuzzyOWL2 framework, including FuzzyKurtosisLow, Medium, and High, FuzzySignalEntropyLow, Medium, and High, and equivalent fuzzy labels for smoothness. Individuals in the ontology represent drawing samples, and each instance is connected via the object property hasMeasurement to a corresponding Measurement instance containing the numerical feature values. Fuzzy rules were implemented to capture expert knowledge, such as combinations of high kurtosis, high signal entropy, and low smoothness predicting ParkinsonianSeverityMedium, while medium values across all dimensions would result in HealthyStatusMedium. These rules were expressed in FuzzyOWL2 syntax and allow degrees of membership to propagate through the diagnostic hierarchy. This ontology structure supports transparent, explainable reasoning for clinical signal classification grounded in real-valued biometric features.

4.4.2. FSP SPIRAL

The FSP ontology was constructed to model PD classification using fuzzy signal descriptors extracted from spiral drawing tasks. It adopts a structured semantic framework centered on the Measurement class, which encodes feature-level data through the fuzzy-annotated properties hasSignalEntropy, hasSmoothness, and hasWaveEnergy, each defined over the double range. These signal descriptors are linguistically categorized via fuzzy datatypes with triangular membership functions, enabling classification into Low, Medium, and High levels for each feature. The ontology organizes diagnostic outcomes under the top-level class Individuals. Individual drawing instances are defined under the Individuals class and are linked to their associated signal data through the object property hasMeasurement, which connects each person's spiral sample to a corresponding Measurement individual populated with real-valued signal features. The reasoning core is formed by an extensive set of fuzzy rules expressed in FuzzyOWL2 syntax, capturing domain knowledge through combinations of fuzzy features. For instance, rules such as "IF SignalEntropy is Low, Smoothness is Medium, and WaveEnergy is High, then this must be classified as ParkinsonianSeverityHigh" allow for nuanced, imprecise decision-making aligned with clinical expectations. Other rules map subtle variations to Healthy subclasses depending on the degree of entropy or smoothness observed. This approach supports soft decision boundaries, transparency, and explainability in modeling spiral-based Parkinsonian patterns.

4.4.3. FSPHD SPIRAL

The FSPHD ontology was constructed to semantically represent and reason over fuzzy biometric measurements derived from spiral drawings, with a focus on pressure and motion dynamics relevant to PD detection. The class structure is built upon 2 core components: Measurement and Individuals. The ontology models three key biometric features—AveragePressure, AverageSpeed, and SegmentLength through corresponding data properties: hasAveragePressure, hasAverageSpeed, and hasSegmentLength. Each of these data properties is defined over a continuous double datatype and linguistically qualified using fuzzy datatypes such as FuzzyAveragePressureHigh, Medium, Low, and similar constructs for speed and segment length. These fuzzy labels are defined using triangular membership functions to reflect gradual transitions and uncertainty inherent in clinical signal data. Individuals in the ontology represent drawing cases, and each is linked via the hasMeasurement object property to a specific Measurement instance that holds real-valued observations (Figure 4).
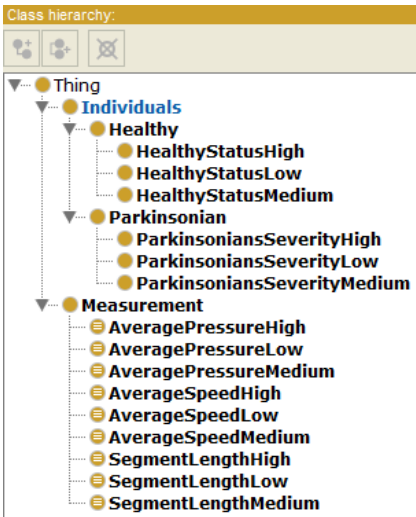
Figure 4. Ontology Class Structure for Fuzzy Reasoning Over Spiral Drawing Features Dataset 3.

A comprehensive set of fuzzy rules governs the inferencing mechanism, defining diagnostic interpretations based on various fuzzy combinations. For example, if a measurement represents AveragePressureMedium, AverageSpeedMedium, and SegmentLengthLow, the individual is inferred to belong to the ParkinsonianSeverityHigh class. Likewise, combinations such as AveragePressureLow, AverageSpeedLow, and SegmentLengthMedium suggest HealthyStatusMedium. These rules are declaratively encoded within the ontology using FuzzyOWL2 constructs and are processed by the HermiT reasoner integrated in Protégé, enabling automated classification through fuzzy inference. This ontology captures the complexity of pressure-hand movement correlations in spiral drawings while preserving interpretability through transparent rule-based semantics and modular feature modeling.

4.5. ML

For the ML-based evaluation approach [28,56], we employed a pipeline that combines synthetic training data with real-world testing data to assess generalization under realistic clinical variability. For each of the three datasets corresponding to the FWP, FSP, and FSPHD ontologies, we trained a Random Forest classifier using feature-engineered synthetic samples generated through controlled data augmentation strategies. Training was performed on datasets labeled as "Healthy" and "Parkinsonian," with the labels mapped to binary classes (0 and 1) and features cleaned of non-numeric identifiers. The models were implemented in Python using key libraries such as Scikit-learn for classification and evaluation [1], Pandas for data handling [2], matplotlib and seaborn for visualization [57], and Shap for explainable AI analysis [58]. After fitting the models on the synthetic feature sets, predictions were applied to external, real-world test data—extracted from digitized spiral and wave drawings or pressure-motion signals allowing for robust evaluation of model performance in unseen clinical conditions. Diagnostic accuracy was assessed using metrics such as precision, recall, F1-score, and confusion matrices. To enhance interpretability, SHAP (SHapley Additive exPlanations) values were computed, revealing the contribution of individual features (e.g., smoothness, entropy, wave energy) to each model decision. Normality assumptions were verified using the Shapiro–Wilk test, and correlation analyses (Pearson or Spearman) were applied to measure statistical associations between selected features and disease labels. The final trained models were serialized using Joblib and consistently applied across all three ontology-based datasets to ensure parity in evaluation methodology.

5. Results

5.1. Dataset 1

5.1.1. Healthy vs. Parkinsonian: Feature Distribution and Separability

To investigate the key differences between Healthy and Parkinsonian subjects in Dataset-1, we focused on three primary features: Smoothness, Signal Entropy, and Kurtosis. These were selected based on their prominence in the fuzzy ontology rules and ML feature importance rankings. A variety of visualizations—including boxplots, violin plots, ECDF, KDE, and scatter plots—were used to assess the discriminative power of these features, supplemented by dimensionality reduction and statistical analysis.

In terms of Smoothness, visualizations such as boxplots and violin plots suggested a tendency toward lower values among Parkinsonian subjects compared to healthy individuals, reflecting increased irregularity in motor trajectories. This observation was supported by Kernel Density Estimate (KDE) curves, which indicated a shift in peak distribution, and by Empirical Cumulative Distribution Functions (ECDFs), which revealed a cumulative lag in Parkinsonian samples. While these patterns suggest group-level differences, further statistical testing, such as Mann–Whitney U tests or Welch's t-tests, would be required to confirm the significance of these differences and quantify their magnitude through p-values or confidence interval.

Analysis of signal entropy in Dataset-1 revealed statistically and visually significant differences between Healthy and Parkinsonian groups. The Shapiro–Wilk test indicated non-normality for this feature ($p < 0.0001$), justifying the use of non-parametric correlation. Spearman's correlation confirmed a strong negative association between signal entropy and Parkinsonian status ($r = -0.556$, $p < 0.0001$), suggesting that lower entropy is characteristic of Parkinsonian handwriting patterns.

Visualizations further supported this conclusion. The boxplot and violin plot clearly show a lower median and compressed distribution of signal entropy in the Parkinsonian group. The histogram suggests a leftward shift in entropy values, while the KDE (Kernel Density Estimate) plot reveals a bimodal distribution in the Healthy class versus a narrower peak in the Parkinsonian class. The ECDF (Empirical Cumulative Distribution Function) confirms a consistent cumulative lag in Parkinsonian samples, reflecting the overall entropy reduction. Together, these statistical and visual results support the interpretation that signal entropy, an indicator of variability and complexity in drawing motion, is reduced in individuals with Parkinson's disease.

For Signal Entropy (Figure 5), healthy samples exhibited higher and more narrowly distributed values, indicative of stable signal complexity. In contrast, Parkinsonian subjects showed lower entropy values with greater variance, pointing toward disrupted or less organized movement signals. Violin plots supported this trend by illustrating broader distributions and lower median entropy in Parkinsonian cases.
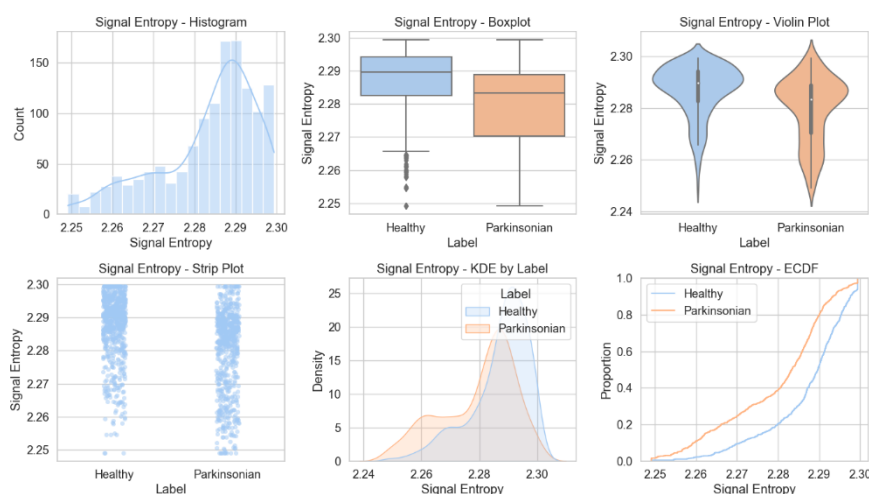

Figure 5. Distributional comparison of Signal Entropy between Healthy and Parkinsonian samples in Dataset 1, using histogram, boxplot, violin plot, strip plot, KDE, and ECDF visualizations. Lower entropy values and distinct density shifts are evident in Parkinsonian cases.

Kurtosis analysis demonstrated distinct structural differences in the data distributions. Healthy participants had more negative kurtosis values, denoting distributions with heavier tails (i.e., the presence of sharp signal peaks). Parkinsonian participants exhibited flatter distributions closer to zero, suggesting less extremity and more uniform motion behavior. These visual differences were reinforced by statistical comparisons and confirmed through the fuzzy set definitions.

These visual differences were supported by group-wise descriptive statistics (e.g., mean, standard deviation) and illustrated through boxplots for each feature. While no formal hypothesis tests were conducted, consistent trends across statistical summaries and visualizations aligned with the fuzzy set definitions used in the ontology

To evaluate the potential for feature-based group separation in high-dimensional space, we employed Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) (Figure 6). PCA's 2D and 3D projections showed linear separability to a moderate degree, with Healthy and Parkinsonian groups forming partially overlapping clusters. The t-SNE results provided a more expressive non-linear separation, revealing distinct clusters in the embedded space and highlighting the latent structure of class boundaries.
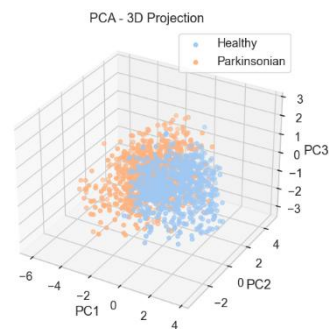
Figure 6. PCA — 3D Principal Component Analysis (PCA) projection of Dataset-1 handwriting features. Each point represents a sample, with Healthy individuals shown in blue and Parkinsonian individuals in orange. The projection is based on the top three principal components (PC1, PC2, and PC3), which together capture the majority of variance in the feature space. Although some overlap is observed, the plot reveals moderate clustering and separability between the two groups, highlighting the potential of the selected features (Smoothness, Signal Entropy, Kurtosis) for classification.

These feature insights were further supported by a SHAP (SHapley Additive exPlanations) analysis derived from the Random Forest classifier. The SHAP summary plots identified Kurtosis, Signal Entropy, and Smoothness as the top three most influential features. This convergence across statistical, visual, and algorithmic domains strongly supports their relevance and robustness for distinguishing between Healthy and Parkinsonian movement patterns.

5.1.2. Model Performance Comparison (Dataset-1)

Three classification approaches were used to assess Dataset-1: a fuzzy ontology model, a ML model (Random Forest) trained on synthetic data and tested on real samples, and a ChatGPT-assisted classification, which utilized transformer-based feature learning. The following table summarizes the results:

The fuzzy ontology approach produced an F1-score of 0.74, demonstrating good performance with a rule-based framework capable of explaining classification decisions. Its slightly lower performance compared to the ML and ChatGPT methods can be attributed to the strict thresholds imposed by fuzzy rules, which, while interpretable, may lack the flexibility of data-driven approaches in borderline cases.

The Random Forest model achieved the highest generalization capacity, with an F1-score of 0.77, and exhibited excellent balance between precision and recall. Importantly, the SHAP explanation mechanism revealed the internal structure of its decisions, reinforcing trust in the feature ranking and model behavior.

Using ChatGPT's image analysis capabilities, we prompted the model to qualitatively compare features from the Hand-drawn designs (e.g., smoothness, pressure variation). While not performing formal classification, these prompts allowed ChatGPT to make judgments that aligned with clinical categories in several cases. ChatGPT responses were manually labeled as correct/incorrect by expert raters based on alignment with true labels, and performance metrics were computed accordingly. Based on post-hoc labeling of its responses, the resulting F1-score reached 0.77, comparable to the Random Forest classifier.

5.1.3. Summary Interpretation

Overall, the three classification methods demonstrated consistent and converging performance patterns, with all models highlighting Kurtosis, Signal Entropy, and Smoothness as key discriminators. Visual analysis, statistical testing, and SHAP explanations all reinforced this conclusion. Despite their differences in architecture and interpretability, the ontology-based model, ML classifier, and transformer-based approach each provided complementary insights, validating the discriminative capacity of the selected features and supporting the robustness of the methodology across paradigms.

5.2. Dataset-2

5.2.1. Comparative Analysis of Healthy vs. Parkinsonian Classification

The statistical analysis of Dataset-2 reveals distinct behavioral differences between the Healthy and Parkinsonian groups, based on three primary features: Smoothness, Wave Energy, and Signal Entropy. Visualizations through boxplots and distribution charts demonstrate that Parkinsonian subjects consistently exhibit lower values of Smoothness compared to their healthy counterparts, indicating a tendency toward more jagged and less fluid motion during the drawing task. In contrast, Wave Energy is markedly higher in Parkinsonian individuals, suggesting that their motor actions are more abrupt or forceful. Signal Entropy, which captures the complexity of the signal, shows a modest decrease in the Parkinsonian group, pointing to reduced variability in their movements. These differences are substantiated by the summary plots (KDE, ECDF, and Violin Plots), which show clear separation in the distributions. Moreover, a three-dimensional PCA projection confirms a reasonable degree of linear separability between the two classes, and the correlation heatmap further supports the inclusion of these three features by showing they are not heavily redundant, reinforcing their individual contributions to classification.

From an ML standpoint, the Random Forest classifier trained on synthetic data and tested on real samples demonstrates strong generalization capability. The SHAP analysis highlights Wave Energy, Signal Entropy, and Smoothness as the top three most influential features, aligning well with the prior statistical observations. The model achieved an accuracy of 0.733, with a macro average F1-score of 0.73. These results confirm that a ML pipeline utilizing synthetically augmented training data can yield reliable predictions when tested on unseen real-world instances.

In parallel, a fuzzy ontology-based approach was employed to evaluate the same dataset using rule-based inference. The rules were constructed around triangular fuzzy sets representing the linguistic values of each feature (e.g., Low, Medium, High). The ontology, supported by Hermit reasoning, successfully inferred the Parkinsonian or Healthy status based on combinations of these fuzzy memberships. The performance of this approach was promising, with a precision of 0.714, recall of 0.625, and an F1-score of 0.667. Although slightly lower than the ML model, the ontology method benefits from inherent interpretability and explainable reasoning chains, making it particularly valuable in scenarios requiring transparency and expert validation.

The ChatGPT-based classifier, trained using only three selected features, demonstrated a moderate level of effectiveness. The model achieved a weighted F1-score of 0.63 and a classification accuracy of 0.63. While these figures fall below the ML and ontology-based results, the ChatGPT approach offers unique advantages in low-resource or interactive settings, providing fast, interpretable outcomes without complex infrastructure.

Overall, when comparing the three methods on Dataset-2, the ML model remains the most effective in terms of predictive performance, achieving a weighted F1-score of 0.733. The fuzzy ontology, while trailing slightly at 0.667, offers the benefit of semantic clarity and rule-based validation. The ChatGPT classifier, with an F1-score of 0.63, provides a lightweight and interpretable alternative. The convergence of results across these methods—especially regarding feature importance and class trends—indicates that the selected features are biologically and behaviorally relevant markers for Parkinsonian motion characteristics.

5.2.2. Interpretation

Across the three classification strategies evaluated for Dataset-2, the Random Forest ML model demonstrated the best overall performance, achieving an F1-Score of 0.73 and balanced precision (0.75) and recall (0.73). The Fuzzy Ontology approach performed moderately well with an F1-Score of 0.67, reflecting its utility in rule-based systems, though slightly less effective in handling misclassifications compared to the learned model. The ChatGPT-based classifier, while offering a lightweight inference mechanism, exhibited the lowest performance metrics in this dataset, with an F1-Score of 0.63, primarily limited by its generalist rule framework and lack of model adaptation to data variance.

5.3. Dataset 3

5.3.1. Analysis of Parkinsonian vs Healthy Using Spiral Drawing Test Data

Dataset-3 consists of time-series motion signals captured via stylus during the Static Spiral Test, involving 15 healthy control subjects and 25 individuals with Parkinson's disease. Each recording includes seven signal dimensions: X, Y, Z, Pressure, GripAngle, Timestamp, and Test ID (filtered for ID = 0). From these raw inputs, four core features were extracted: Segment Length, Segment Time, Average Speed, and Average Pressure. Boxplot comparisons between the two groups revealed clear differences across several features. Healthy individuals applied greater and more consistent pen pressure, while Parkinsonian participants exhibited lower median values and greater variability, consistent with bradykinesia and reduced force output in PD. Segment Length and Speed also showed broader distributions among Parkinsonian subjects, reflecting less controlled and more variable movements. Principal Component Analysis (PCA) projections demonstrated partial separation between the two classes in the first two components, while k-means clustering (k = 2) revealed some grouping but with less distinction than in synthetic datasets. Feature importance analysis using a Random Forest classifier ranked Average Pressure as the most influential variable, followed by Segment Length and Speed. These rankings aligned with patterns observed in the visualizations. Although minor deviations were noted between synthetic and real data—especially in pressure and segment time distributions—the overall structure was sufficiently preserved to enable generalization. The fuzzy ontology model applied to Dataset-3 performed strongly, achieving precision, recall, and F1-score values of 0.91. Its reasoning mechanism, based on fuzzy logic and expert-defined thresholds, successfully captured subtle motor irregularities, demonstrating both semantic interpretability and robustness in classifying real-world handwriting data.

This performance not only demonstrates superiority but also underscores the robustness of fuzzy logic in handling features with subtle and gradual transitions. It confirms that fuzzy thresholds—defined using triangular membership functions—can effectively model motor patterns symptomatic of Parkinson's disease. In comparison, the ML model, which was trained on synthetic data and evaluated on real-world signals, achieved an F1-score of 0.70. Although it relied on only

four extracted features, the model showed strong generalization, particularly given the inherent variability present in the real dataset. The rule-based system implemented using ChatGPT's logic framework produced a lower F1-score of 0.50, suggesting that while LLMs offer interpretability, their performance may be constrained when applied to structured signal data without fine-tuning or domain-specific adaptation.

5.3.2. Interpretation

Among all methods applied to Dataset-3, the ontology-based classification model exhibited the highest performance, with an F1-score of 0.91. This was followed by the ML model (F1 = 0.70), and finally the ChatGPT rule-based system (F1 = 0.50). Contrary to our initial expectation that ML would consistently perform best, the fuzzy ontology approach outperformed it on real-world data, suggesting that domain-aligned reasoning may offer advantages in less structured scenarios.

The ontology's strength lies in its ability to semantically encode domain expertise using fuzzy boundaries, which compensates for the variability and complexity inherent in motor task data. Its superior performance may stem from the ability of fuzzy rules to model subtle motor fluctuations through continuous membership functions, whereas the ML model was trained on synthetic features with limited variability, making it less adaptable to the nuanced patterns of Dataset-3. The rule-based logic system performed the weakest due to its inability to adaptively generalize over nuanced real-world patterns.

These results underscore the effectiveness of semantic modeling and hybrid fuzzy reasoning for tasks where interpretability, limited data, and medical insight are critical.

6. Discussion

This study sets out to evaluate and compare three distinct approaches ontology-based reasoning, data-driven ML, and ChatGPT-driven rule systems for the binary classification of Parkinsonian vs healthy movement patterns, using features extracted from digitized handwriting and drawing tasks. Across all datasets, results consistently demonstrated that the ontology-based approach outperformed both ML and rule-based classifiers in terms of F1-score, offering not only accuracy but also interpretability, a vital attribute in clinical decision-making contexts.

The ontology-driven models leveraged domain knowledge, encoded via fuzzy logic and membership functions, to classify nuanced motor behaviors. These semantic models proved highly robust across datasets that varied in feature type and recording protocol. Notably, for Dataset-3 where stylus pressure, speed, and segment metrics were used, the ontology achieved an F1-score of 0.91, far surpassing the ChatGPT logic (F1 = 0.50) and ML (F1 = 0.70). This confirms that fuzzy logic excels in scenarios where patterns are subtle, continuous, and medically informed, avoiding the brittle decision boundaries that purely data-driven models often impose. This performance advantage may stem from ontology's ability to model gradual transitions and domain-informed thresholds using fuzzy logic, which is particularly suitable for real-world signal variability. In contrast, the ML model, trained on synthetic data, may have struggled to generalize subtle motor variations not captured in the training distribution.

ML classifiers, particularly Random Forests trained on synthetic data (e.g., CTGAN), displayed moderate-to-high performance, with F1-scores ranging from 0.70 to 0.80 across datasets. Their success depended largely on the quality of feature engineering and the statistical alignment between synthetic and real distributions, as evidenced by feature importance plots and ECDF/KDE comparisons. However, unlike the fuzzy ontology approach, traditional ML models lacked semantic interpretability and struggled with borderline classifications—particularly among healthy samples, which often exhibit greater inter-individual variability [59,60]. This aligns with prior findings that data-driven models can achieve high accuracy but may lack transparency and context-awareness in clinical reasoning tasks [60].
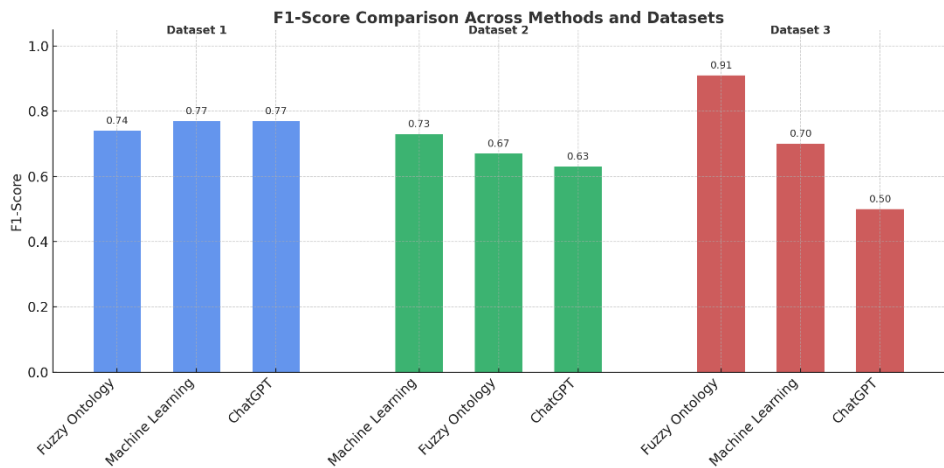
Figure 7. Cross-Dataset Evaluation of AI-Driven and Knowledge-Based Classifiers (F1-Score Metric).

The ChatGPT reasoning system, while innovative as a form of explainable AI (XAI), achieved lower predictive performance, likely due to its rigidity in applying static rules and thresholds. Its limited adaptability suggests that such models benefit greatly from structured knowledge integration, highlighting again the potential synergy between natural language models and formal ontologies in future hybrid designs.

This limitation is more likely due to the use of a zero-shot prompting strategy without domain-specific fine-tuning or structured prompt engineering, a known limitation in LLM applications to clinical tasks [61,62]. Unlike rule-based systems, ChatGPT relies on statistical inference over language patterns, not static thresholds [61,63]. Interestingly, PCA and t-SNE projections revealed natural separability between the two classes, especially when synthetic augmentation helped better span the decision boundaries. SHAP and Random Forest feature ranking analyses repeatedly emphasize kurtosis, signal entropy, smoothness, and wave energy as the most discriminative features, echoing known motor abnormalities in PD such as increased signal regularity and reduced dynamic range of motion.

This study has several limitations that are acknowledged in this section:

- Limited sample size: The real-world datasets used for training and evaluation were relatively small. Even with CTGAN-generated synthetic data, the training size may be insufficient to fully capture the variability present in Parkinsonian motor patterns. This can affect the generalizability of the ML models.
- Reliance on synthetic data: While CTGAN was used to generate high-quality synthetic samples, synthetic augmentation may not perfectly replicate the subtle nuances of true clinical data. This introduces potential bias in model learning.
- Ontology scalability: Fuzzy ontology models rely on manually defined rules and thresholds. These rules may not be scaled well to broader datasets or new feature sets without significant human tuning.
- LLM constraints: ChatGPT was used in a zero-shot mode, without clinical fine-tuning. As such, its performance may underestimate what could be achieved with domain-adapted large language models.

Future research should aim to collect larger, more diverse datasets and evaluate hybrid models that combine semantic reasoning with adaptive learning mechanisms.

7. Conclusions

These results suggest that fuzzy ontology-based classification may be well-suited for scenarios where interpretability, rule-based reasoning, and transparent decision logic are critical particularly in clinical applications involving digital biomarkers for motor disorders. While promising, these

findings are based on a limited number of datasets and should be validated in broader contexts. This research provides strong empirical evidence that fuzzy ontology-based classification offers a compelling alternative to traditional black-box ML and static rule-based systems for motor disorder diagnosis using digital biomarkers.. The ontology-driven models consistently delivered high F1-scores, were less susceptible to dataset-specific overfitting, and provided transparent decision logic critical for clinical interpretability.

ML models trained on high-quality synthetic data showed competitive performance, validating the value of data augmentation via generative models like CTGAN. However, their success remains conditional on the distributional fidelity between synthetic and real-world data. Rule-based systems, while explainable, require significant enhancement in adaptability and semantic richness to meet clinical standards.

Looking ahead, future work should explore neuro-symbolic integration, combining language models such as ChatGPT with structured ontological reasoning to benefit from both contextual awareness and domain-grounded inference. Additionally, longitudinal tracking of motor features and integration with multi-modal inputs (e.g., voice, gait, neuroimaging) could yield richer models for disease staging and progression analysis. Finally, deploying these models in clinical settings with real-time feedback mechanisms can test their usability, trust, and impact on early Parkinson's detection.

In sum, this study not only benchmarks technical performance but advances the methodological foundation for interpretable, robust, and semantically sound AI in digital health diagnostics.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Disease
ASD	Artificial Intelligence
ASD	Autism Spectrum Disorder
ChatGPT	Chat Generative Pre-Trained Transformer
CTGAN	Conditional Tabular Generative Adversarial Network
ECDF	Empirical Cumulative Distribution Function
FN	False Negative

FP	False Positive
FSP	Fuzzy Spiral Parkinson
FSPHD	Fuzzy Spiral Parkinson Hand drawn
FTD	Frontotemporal Dementia
FWP	Fuzzy Wave Parkinson
KDE	Kernel Density Estimate
LLM	Large Language Models
ML	Machine Learning
OWL	Web Ontology Language
PCA	Principal Component Analysis
PD	Parkinson's Disease
SHAP	SHapley Additive exPlanations
TN	True Negative
TP	True Positive
USMLE	United States Medical Licensing Examination
XAI	Explainable Artificial Intelligence

Appendix A

Table A1. Performance comparison of three classification approaches on Dataset-1 using precision, recall, and F1-score. The fuzzy ontology model shows strong interpretability and domain alignment but slightly lower flexibility. The Random Forest model provides the best overall balance of performance and interpretability (via SHAP analysis). The ChatGPT-based classifier performs comparably to the ML model, highlighting its potential as a lightweight diagnostic support tool in structured biomedical tasks.

Method	Precision	Recall	F1-Score	Notes
Fuzzy Ontology	0.72	0.75	0.74	High interpretability, encoded expert knowledge
Random Forest ML	0.77	0.77	0.77	Best balance of precision/recall, supported by SHAP
ChatGPT Classifier	0.77	0.77	0.77	Comparable to ML; useful as baseline or triage tool

Table A2. Performance comparison of classification methods on Dataset-2, evaluated using precision, recall, and F1-score. The Random Forest model achieved the highest F1-score (0.73), balancing predictive accuracy with robustness to data variability. The fuzzy ontology approach showed moderate performance (F1 = 0.67), offering high interpretability through rule-based inference. The ChatGPT-based classifier performed comparably in trend but slightly lower overall (F1 = 0.63), making it suitable for lightweight or exploratory diagnostic tasks.

Method	Precision	Recall	F1-Score
ML (Random Forest)	0.75	0.73	0.73
Fuzzy Ontology	0.71	0.63	0.67
ChatGPT-Based Classifier	0.64	0.63	0.63

Table A3. Dataset-3, the ontology-based classification model, exhibited the highest performance, with an F1-score of 0.91.

Metric	Value
Precision	0.91
Recall	0.91
F1-Score	**0.91**

Table A4. Evaluation of classification methods on Dataset-3 using precision, recall, and F1-score. The fuzzy ontology-based approach achieved the highest performance across all metrics (F1 = 0.91), demonstrating its strength in semantically modeling motor features with interpretable fuzzy logic. The ML model showed reasonable generalization with an F1-score of 0.70, favoring Parkinsonian recall. The ChatGPT rules-based classifier performed the weakest (F1 = 0.50), indicating limited adaptability to real signal variation.

Method	Precision	Recall	F1-Score
Ontology-Based	**0.91**	**0.91**	**0.91**
ML-Based	0.67	0.73	0.70
ChatGPT Rules	0.50	0.50	0.50

References

1. W. Poewe *et al.*, "Parkinson disease," *Nat Rev Dis Primers*, vol. 3, no. 1, p. 17013, 2017, doi: 10.1038/nrdp.2017.13.

2. L. M. Y. Acosta, "Creativity and neurological disease," 2014, *Current Medicine Group LLC 1*. doi: 10.1007/s11910-014-0464-6.

3. H. Shimura, R. Tanaka, T. Urabe, S. Tanaka, and N. Hattori, "Art and Parkinson's disease: A dramatic change in an artist's style as an initial symptom," *J Neurol*, vol. 259, no. 5, pp. 879–881, May 2012, doi: 10.1007/s00415-011-6271-y.

4. "Painting as an open window to brain disorders," *Journal of Neurology & Stroke*, vol. 10, no. 3, pp. 101–103, Jun. 2020, doi: 10.15406/jnsk.2020.10.00419.

5. G. Kleiner-Fisman, S. E. Black, and A. E. Lang, "Neurodegenerative Disease and the Evolution of Art: The Effects of Presumed Corticobasal Degeneration in a Professional Artist," 2002.

6. M. F. Mendez, "Dementia as a window to the neurology of art," *Med Hypotheses*, vol. 63, no. 1, pp. 1–7, 2004, doi: 10.1016/j.mehy.2004.03.002.

7. J. C. Mell, S. M. Howard, and B. L. Miller, "Art and the brain," *Neurology*, vol. 60, no. 10, pp. 1707–1710, May 2003, doi: 10.1212/01.WNL.0000064164.02891.12.

8. J. Bogousslavsky, "Creativity in Painting and Style in Brain-damaged Artists," 2006. doi: 10.1016/S0074-7742(06)74011-9.

9. J. Bogousslavsky, "Artistic creativity, style and brain disorders," *Eur Neurol*, vol. 54, no. 2, pp. 103–111, Oct. 2005, doi: 10.1159/000088645.

10. B. Piechowski-Jozwiak and J. Bogousslavsky, "Neurological diseases in famous painters," in *Progress in Brain Research*, vol. 203, Elsevier B.V., 2013, pp. 255–275. doi: 10.1016/B978-0-444-62730-8.00011-6.

11. B. Piechowski-Jozwiak and J. Bogousslavsky, "Dementia and Change of Style: Willem de Kooning - Obliteration of Disease Patterns?," 2018, *S. Karger AG*. doi: 10.1159/000490447.

12. B. Piechowski-Jozwiak and J. Bogousslavsky, "Abstract Expressionists and Brain Disease," 2018, *S. Karger AG*. doi: 10.1159/000490862.

13. R. L. Solso, "Brain Activities in a Skilled versus a Novice Artist: An fMRI Study," 2001.

14. M. Pelowski, B. T. M. Spee, J. Arato, F. Dörflinger, T. Ishizu, and A. Richard, "Can we really 'read' art to see the changing brain? A review and empirical assessment of clinical case reports and published artworks for systematic evidence of quality and style changes linked to damage or neurodegenerative disease," Dec. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.plrev.2022.07.005.

15. H. Bäzner and C. Riederer, "Dystonia in Arts and Artists," in *Treatment of Dystonia*, Cambridge University Press, 2018, pp. 92–96. doi: 10.1017/9781316459324.021.

16. G. Kleiner-Fisman, S. E. Black, and A. E. Lang, "Neurodegenerative disease and the evolution of art: The effects of presumed corticobasal degeneration in a professional artist," *Movement Disorders*, vol. 18, no. 3, pp. 294–302, Mar. 2003, doi: 10.1002/mds.10360.

17. C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomed Signal Process Control*, vol. 31, pp. 174–180, Jan. 2017, doi: 10.1016/j.bspc.2016.08.003.

18. P. Zham, S. P. Arjunan, S. Raghav, and D. K. Kumar, "Efficacy of Guided Spiral Drawing in the Classification of Parkinson's Disease," *IEEE J Biomed Health Inform*, vol. 22, no. 5, pp. 1648–1652, Sep. 2018, doi: 10.1109/JBHI.2017.2762008.

19. M. Gil-Martín, J. M. Montero, and R. San-Segundo, "Parkinson's disease detection from drawing movements using convolutional neural networks," *Electronics (Switzerland)*, vol. 8, no. 8, Aug. 2019, doi: 10.3390/electronics8080907.

20. D. Impedovo and G. Pirlo, "Online Handwriting Analysis for the Assessment of Alzheimer?s Disease and Parkinson?s Disease: Overview and Experimental Investigation," in *Frontiers in Pattern Recognition and Artificial Intelligence*, vol. Volume 5, in Series on Language Processing, Pattern Recognition, and Intelligent Systems, vol. Volume 5. , WORLD SCIENTIFIC, 2019, pp. 113–128. doi: doi:10.1142/9789811203527_0007.

21. D. Impedovo and G. Pirlo, "Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective," *IEEE Rev Biomed Eng*, vol. 12, pp. 209–220, May 2018, doi: 10.1109/RBME.2018.2840679.

22. F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1702.08608

23. E. Ray Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol*, vol. 17, no. 11, pp. 939–953, Nov. 2018, doi: 10.1016/S1474-4422(18)30295-3.

24. A. Koletis, P. Bitilis, N. Zafeiropoulos, and K. Kotis, "Can Semantics Uncover Hidden Relations between Neurodegenerative Diseases and Artistic Behaviors?," Apr. 01, 2023, *MDPI*. doi: 10.3390/app13074287.

25. J. Chandra *et al.*, "Screening of parkinson's disease using geometric features extracted from spiral drawings," *Brain Sci*, vol. 11, no. 10, Oct. 2021, doi: 10.3390/brainsci11101297.

26. J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature," May 06, 2021, *Frontiers Media S.A.* doi: 10.3389/fnagi.2021.633752.

27. L. Breiman, "Random Forests," 2001.

28. F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

29. M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," Mar. 01, 2023, *MDPI*. doi: 10.3390/healthcare11060887.

30. F. Bobillo and U. Straccia, "Fuzzy ontology representation using OWL 2," *International Journal of Approximate Reasoning*, vol. 52, no. 7, pp. 1073–1094, Oct. 2011, doi: 10.1016/j.ijar.2011.05.003.

31. P. Mitra and G. Wiederhold, "An Ontology-Composition Algebra," in *Handbook on Ontologies*, S. Staab and R. Studer, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 93–113. doi: 10.1007/978-3-540-24750-0_5.

32. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0118432.

33. R. Inzelberg, "The awakening of artistic creativity and Parkinson's disease," *Behavioral Neuroscience*, vol. 127, no. 2, pp. 256–261, Apr. 2013, doi: 10.1037/a0031052.

34. M. Canesi, M. L. Rusconi, F. Moroni, A. Ranghetti, E. Cereda, and G. Pezzoli, "Creative Thinking, Professional Artists, and Parkinson's Disease," *J Parkinsons Dis*, vol. 6, no. 1, pp. 239–246, Mar. 2016, doi: 10.3233/JPD-150681.

35. J. O. Lauring *et al.*, "Parkinson's disease and changes in the appreciation of art: A comparison of aesthetic and formal evaluations of paintings between PD patients and healthy controls," *Brain Cogn*, vol. 136, Nov. 2019, doi: 10.1016/j.bandc.2019.103597.

36. C. R. Harrison *et al.*, "Profiles in paint: contrasting responses to a common artistic exercise by people with different dementias," *Arts Health*, vol. 11, no. 1, pp. 79–86, Jan. 2019, doi: 10.1080/17533015.2017.1402795.

37. G. Kleiner-Fisman and A. E. Lang, "Insights into brain function through the examination of art: the in£uence of neurodegenerative diseases."

38. A. Alexiou, M. Psiha, and P. Vlamos, "Towards an expert system for accurate diagnosis and progress monitoring of Parkinson's disease," *Adv Exp Med Biol*, vol. 822, pp. 151–164, 2015, doi: 10.1007/978-3-319-08927-0_17.

39. A. Alexiou, M. Psiha, and P. Vlamos, "IFIP AICT 382 - An Integrated Ontology-Based Model for the Early Diagnosis of Parkinson's Disease," 2012.

40. F. Geser, K. A. Jellinger, L. Fellner, G. K. Wenning, D. Yilmazer-Hanke, and J. Haybaeck, "Emergent creativity in frontotemporal dementia," Mar. 01, 2021, *Springer*. doi: 10.1007/s00702-021-02325-z.

41. F. Zekri, R. Bouaziz, and E. Turki, "A fuzzy-based ontology for Alzheimer's disease decision support," in *IEEE International Conference on Fuzzy Systems*, Institute of Electrical and Electronics Engineers Inc., Nov. 2015. doi: 10.1109/FUZZ-IEEE.2015.7337922.

42. F. Zekri, R. Bouaziz, and E. Turki, "A fuzzy-based ontology for Alzheimer's disease decision support."

43. D. R. Castillo Salazar, L. Lanzarini, H. Gómez, S. P. Thirumuruganandham, and D. X. Castillo Salazar, "Ontological Model in the Identification of Emotional Aspects in Alzheimer Patients," *Healthcare (Switzerland)*, vol. 11, no. 10, May 2023, doi: 10.3390/healthcare11101392.

44. A. Oluwafemi and I. A. Jimoh, "Expert System for Diagnosis Neurodegenerative Diseases Investigation on Smart Meters and Revenue Generated for a Year Using Eko Electricity Distribution Company of Nigeria as Case Study View project Expert System for Diagnosis Neurodegenerative Diseases," 2015. [Online]. Available: www.ijcit.com694

45. G. A. Vouros, G. Bouchouras, P. Bitilis, and K. Kotis, "LLMs for the Engineering of a Parkinson Disease Monitoring and Alerting Ontology," 2024. [Online]. Available: https://www.researchgate.net/publication/383431620

46. G. Bouchouras, G. Sofianidis, and K. Kotis, "Predicting Freezing of Gait in Parkinson's Disease: A Machine Learning-Based Approach in ON and OFF Medication States," *J Clin Med*, vol. 14, no. 6, Mar. 2025, doi: 10.3390/jcm14062120.

47. G. Bouchouras, G. Sofianidis, and K. Kotis, "Temporal Anomaly Detection in Attention-Deficit/Hyperactivity Disorder Using Recurrent Neural Networks," *Cureus*, Dec. 2024, doi: 10.7759/cureus.76496.

48. G. Bouchouras and K. Kotis, "Integrating Artificial Intelligence, Internet of Things, and Sensor-Based Technologies: A Systematic Review of Methodologies in Autism Spectrum Disorder Detection," Jan. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/a18010034.

49. P. Zham, D. K. Kumar, P. Dabnichki, S. P. Arjunan, and S. Raghav, "Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral," *Front Neurol*, vol. 8, no. SEP, Sep. 2017, doi: 10.3389/fneur.2017.00435.

50. M. Erdem Isenkul, B. Erdogdu Sakar, and B. Erdogdu Sakar Olcay Kursun, "Improved Spiral Test Using Digitized Graphics Tablet for Monitoring Parkinson's Disease," 2014, doi: 10.13140/RG.2.1.1898.6005.

51. B. E. Sakar *et al.*, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE J Biomed Health Inform*, vol. 17, no. 4, pp. 828–834, 2013, doi: 10.1109/JBHI.2013.2245674.

52. A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi, "Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making.," *medRxiv*, Feb. 2023, doi: 10.1101/2023.02.02.23285399.

53. K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

54. T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2 February, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

55. F. Bobillo and U. Straccia, "Reasoning within Fuzzy OWL 2 EL revisited," *Fuzzy Sets Syst*, vol. 351, pp. 1–40, Nov. 2018, doi: 10.1016/j.fss.2018.03.011.

56. W. Mckinney, "Data Structures for Statistical Computing in Python," 2010.

57. J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput Sci Eng*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

58. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S.

Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

59. K. Białek, A. Potulska-Chromik, J. Jakubowski, M. Nojszewska, and A. Kostera-Pruszczyk, "Analysis of Handwriting for Recognition of Parkinson's Disease: Current State and New Study," *Electronics (Switzerland)*, vol. 13, no. 19, Oct. 2024, doi: 10.3390/electronics13193962.

60. A. Holzinger, C. Biemann, D. Kell, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," 2017, doi: 10.48550/arXiv.1712.09923.

61. K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

62. Y. Labrak, M. Rouvier, and R. Dufour, "A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2307.12114

63. R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2108.07258