Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantifying Data Leakage in Multimodal Clinical Augmentation: A Decomposition Framework and a Leakage-Free Evaluation Protocol for Parkinson's Disease Classification

Mohamed Hedi Djemaa
[1]INU Champollion, CHART, University of Toulouse
Yohann Chasseray
[1]INU Champollion, CHART, University of Toulouse
Rafika Thabet
Institut Supérieur d'Informatique et des Technologies de Communication de Hammam Sousse
Farah Jemili
jmili_farah@yahoo.fr

Institut Supérieur d'Informatique et des Technologies de Communication de Hammam Sousse
Elyes Lamine
[1]INU Champollion, CHART, University of Toulouse

Research Article

Keywords: Multimodal learning, Latent-space augmentation, Class imbalance, Synthetic data generation, Data leakage, Parkinson's disease, Clinical machine learning evaluation.

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10534776/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quantifying Data Leakage in Multimodal Clinical Augmentation: A Decomposition Framework and a Leakage-Free Evaluation Protocol for Parkinson's Disease Classification

Mohamed Hedi Djemaa[1,3], Yohann Chasseray[2], Rafika Thabet[4], Farah Jemili[4] *, Elyes Lamine[2]

[1]INU Champollion, CHART, University of Toulouse, Castres, France. [2]INU Champollion, IRIT-CNRS, University of Toulouse, Castres, France. [3]ENSI, INSIGHT Lab, University of Manouba, Tunis, Tunisia. [4]ISITCom, INSIGHT Lab, University of Sousse, Sousse, Tunisia.

*Corresponding author: jmili_farah@yahoo.fr

Highlights

- Order of embedding and balancing operations does not affect classification AUC in multimodal Parkinson's disease pipelines.
- A novel decomposition framework isolates split-after-augmentation leakage from genuine augmentation gains.
- Leakage inflates reported AUC by up to +0.17 in multiclass Parkinson's disease detection.
- Latent-space GAN augmentation yields near-indistinguishable synthetic patients (C2ST AUC = 0.458) versus raw-space (1.000).
- A released leakage-free protocol offers a reusable clinical-AI audit tool for augmentation claims.

Abstract

Generative augmentation is widely used to correct class imbalance in multimodal clinical machine learning, yet classification gains attributed to it are rarely checked for evaluation leakage, a gap with direct consequences for clinical trust in AI-assisted diagnosis. We ask whether the order in which heterogeneous modalities are embedded and rebalanced with synthetic samples affects downstream Parkinson's disease (PD) classification, and we introduce a decomposition framework that separates any apparent gain into a baseline term, a genuine augmentation effect, and a split-after-augmentation leakage term. Using a multimodal PD dataset combining wearable sensor time-series, clinical scales, and a non-motor symptom questionnaire, we compare Embed-Then-Balance (ETB) and Balance-Then-Embed (BTE) pipelines under a leakage-free 5-fold × 5-seed cross-validation protocol, and assess synthetic-sample fidelity with a classifier two-sample test, the Wasserstein distance, and the distance-to-closest-record. Classification performance is statistically indistinguishable across pipelines and no augmentation method improves on the unaugmented baseline (best AUC 0.838-0.842 across configurations). The two orderings diverge sharply in generative fidelity: raw-space balancing produces off-manifold samples trivially separable from real data (C2ST AUC = 1.000), whereas latent-space balancing produces samples statistically indistinguishable from real ones (C2ST AUC = 0.458). Gains commonly attributed to augmentation are, to a large extent, an artifact of evaluation leakage, contributing up to +0.17 AUC in multiclass settings. For clinical machine learning, this means that leakage-free evaluation is a prerequisite, not an optional refinement, before an augmentation strategy can be trusted for deployment in a diagnostic-support pipeline. The decomposition framework and protocol we release apply to any clinical machine learning study combining representation learning with synthetic-data augmentation.

Keywords

Multimodal learning; Latent-space augmentation; Class imbalance; Synthetic data

generation; Data leakage; Parkinson's disease; Clinical machine learning evaluation.

1. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder and the fastest growing in terms of disability burden, in a context where neurological disorders collectively affect 3.40 billion people worldwide [1]. Clinical diagnosis of PD remains imperfect, with reported accuracy between 73% and 84% [2], and symptom overlap with other movement disorders complicates early-stage identification. Although PD is primarily characterized by motor symptoms (tremor, rigidity, bradykinesia, postural instability), it is increasingly recognized as a multisystem disorder involving cognitive, autonomic, and sleep-related manifestations [3,4]. This heterogeneity motivates multimodal data-driven approaches that integrate complementary signals to support earlier and more reliable detection [5].

From a clinical-workflow perspective, the practical promise of multimodal AI for PD is a screening or triage tool that combines low-burden wearable and questionnaire data to flag patients who warrant specialist neurological assessment. Realizing that promise safely requires that reported performance gains be genuine rather than artifacts of how a study was evaluated: a decision-support tool validated under a leaky protocol risks overstating its benefit once deployed on truly unseen patients, with consequences for both clinical trust and patient safety. This concern motivates the present study.

Multimodal machine learning addresses the detection need by combining sensor data, clinical scales, neuroimaging, voice, and patient reported outcomes, consistently outperforming unimodal baselines for PD detection [6-8]. However, three obstacles limit clinical translation. First, patient cohort mismatches and temporal misalignment yield incomplete multimodal profiles. Second, fusion mechanisms struggle to preserve cross-modal semantics when combining high-dimensional sources (e.g., sensor time-series) with low-dimensional ones (e.g., questionnaires). Third, severe class imbalance at modality intersections biases learning toward majority classes [9]. Generative augmentation, particularly through Generative Adversarial Networks (GANs), has been proposed to mitigate scarcity and imbalance [10,11].

A common pipeline in multimodal clinical learning combines two operations: embedding (via autoencoders) to reduce heterogeneity, and balancing (via SMOTE or GANs) to address class imbalance. While latent-space augmentation has been explored in unimodal settings, such as imbalanced image classification [12] and ECG beat classification [13], the relative ordering of embedding and balancing operations in heterogeneous multimodal clinical pipelines has received little systematic attention, and the methodological risk of split-after-augmentation leakage has received even less. Practitioners typically apply augmentation either before embedding (in raw input space) or after embedding (in latent-space). Yet these two orderings make fundamentally different assumptions about the geometry of the data: raw-space generation must reproduce the high-dimensional manifold structure directly, while latent-space generation operates on a compact, disentangled representation.

In this work, we systematically compare these two pipeline orderings on heterogeneous multimodal clinical data, with an explicit focus on whether reported gains survive a leakage-free evaluation. We define Embed-Then-Balance (ETB), which projects modalities into a unified latent-space before applying GAN-based augmentation, and Balance-Then-Embed (BTE), which generates synthetic samples in the raw

input space before encoding. The principal contributions of this article are:

- A general, modality-agnostic framework for the ordering of embedding and balancing operations, with a manifold-hypothesis argument for why this order is expected to affect the geometry of the synthetic samples each pipeline produces.
- A leakage-free evaluation on a heterogeneous multimodal clinical dataset. We instantiate both orderings on the PADS Parkinson's disease dataset, which combines wearable sensor time-series, clinical scales, and a non-motor symptom questionnaire, and we evaluate them under a cross-validation protocol that fits every preprocessing and augmentation step on training data only. We further introduce a quantitative decomposition framework that separates the apparent gain into a baseline term, a residual augmentation effect, and a split-after-augmentation leakage term — a methodological audit tool directly applicable to other clinical AI evaluations.
- A generative quality evaluation of both orderings, assessed with a classifier two-sample test, the Wasserstein distance, and the distance-to-closest-record, showing that the two orderings differ markedly in the fidelity of the data they generate even where they do not differ in classification.
- An explicit discussion of the clinical implications of these findings for the design and audit of augmentation-based decision-support pipelines (Section 6.5).

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 describes the experimental setup. Section 5 reports the results. Section 6 discusses limitations, clinical implications, and interpretation. Section 7 concludes.

2. Related Work

To position our contribution, we organize prior work into four areas: multimodal learning for neurodegenerative disease diagnosis, generative augmentation under class imbalance, fusion strategies in clinical pipelines, and prior work on the PADS dataset. Table S1 (Supplementary Material) compares our approach to representative methods across these areas.

2.1 Multimodal Learning for Neurodegenerative Disease Diagnosis

Unimodal approaches have demonstrated effectiveness in extracting disease-specific patterns from individual data sources, including EEG signals for Alzheimer's disease [14] and Parkinson's disease [15], spiral drawings [16], and structural MRI [17]. While these methods achieve high accuracy on their target modality, they cannot exploit the complementary information distributed across the motor, cognitive, and autonomic channels affected by PD.

Multimodal approaches address this limitation by integrating heterogeneous signals. Sensor-based monitoring captures gait and tremor in naturalistic settings [18], while combinations of imaging, handwriting, clinical scales, and patient-reported outcomes consistently outperform unimodal baselines for early PD detection [6,19,20]. The same principle has been validated for Alzheimer's disease, where deep learning models combining structural MRI, genetic markers, and cognitive tests reach diagnostic performance comparable to expert neurologists [21].

2.2 Generative Augmentation under Class Imbalance

Class imbalance and data scarcity are pervasive in clinical ML, and Generative Adversarial Networks (GANs) [22] have emerged as a primary tool for synthetic data generation in healthcare [23]. For multivariate clinical time-series, recurrent and conditional GANs have shown that synthetic samples can preserve

predictive utility while addressing class imbalance [24]. Extensions of this paradigm have been applied to PD-specific tasks: GAN-based augmentation of spiral-drawing images improves classification accuracy when training data are limited [25].

A central methodological question concerns where in the pipeline augmentation should occur. Two strategies dominate practice. The first generates synthetic samples directly in the raw input space, then encodes them downstream; this preserves all original feature dimensions but requires the generator to model the full high-dimensional manifold, which is challenging for heterogeneous clinical data. The second strategy first projects inputs into a compact latent-space via autoencoders, then performs augmentation within this representation. Latent-space augmentation has been explored in unimodal settings: [12] couple a variational autoencoder with a GAN (IDA-GAN) to generate minority-class samples in latent-space for imbalanced image classification, and [13] apply a similar autoencoder-plus-GAN strategy to ECG beat classification, reporting AUC improvements of 2% in both binary and multiclass settings. These works establish the value of latent-space augmentation but remain confined to single-modality data and, notably, do not audit their reported gains for split-after-augmentation leakage. Privacy considerations have also shaped this literature: [26] formalize the trade-off between predictive utility and membership-inference risk for clinical time-series GANs, motivating the inclusion of distance-based privacy metrics in synthetic data evaluation.

2.3 Multimodal Fusion Strategies

Multimodal representation learning seeks to project heterogeneous inputs into a shared latent-space where semantically similar content is encoded by similar vectors [27]. Formally, em = f(x1, ..., xn), where xi ∈ R^di denote modality-specific representations and em ∈ R^dshared is the fused embedding. Three fusion regimes are commonly distinguished. Early fusion concatenates modality-specific embeddings before downstream processing, producing fine-grained joint representations but requiring careful temporal and dimensional alignment; this strategy has been used for PD classification combining handwriting, drawing, and clinical scores [6]. Intermediate fusion introduces attention or cross-modal gating mechanisms that weight modality contributions during processing, capturing non-linear inter-modal dependencies; recent PD work in this category includes cross-modal attention with contrastive learning between MRI and clinical scales [28], semi-supervised speech embedding fusion [29], and smartphone-based fusion of voice, finger tapping, and gait [30]. Late fusion integrates independent modality-specific decisions and is robust to missing modalities, but cannot exploit feature-level cross-modal correlations.

Each fusion regime presents trade-offs between expressiveness, robustness, and computational cost. We adopt early fusion because it provides the cleanest experimental setting for isolating the effect of augmentation ordering, our primary research question. The choice of fusion strategy is orthogonal to the ETB-versus-BTE comparison, which can in principle be applied to any of the three regimes.

2.4 Prior Work on the PADS Dataset

The PADS dataset [31] was introduced together with an initial ML benchmark by [32], who reported balanced accuracies of 91.16% for PD-versus-HC and 72.42% for PD-versus-DD using an integrative pipeline combining classical signal processing and deep learning on the smartwatch sensor stream. More recent work has extended this benchmark: [33] propose MFAM, a multi-scale frequency-aware adversarial multi-instance network that leverages medical priors for PD-

versus-DD severity assessment on PADS. These prior efforts focus on model architecture for signal-based classification, but do not address the orthogonal question of how augmentation strategies should be combined with multimodal embedding when class imbalance is severe, nor whether their reported gains are leakage-free.

Table S1 (Supplementary Material) summarizes the position of our contribution against representative prior work across four criteria: (i) domain and data scope (healthcare, multimodal); (ii) methodological pillars (fusion, embedding, augmentation); (iii) modeling strategy (patient correspondence, multiclass capability); and (iv) evaluation rigor (classification metrics, generative quality metrics, leakage control). Prior work typically addresses a subset of these criteria. Unimodal approaches lack fusion robustness; augmentation-focused studies rarely evaluate multiclass detection performance jointly with generative quality or leakage control; fusion-focused architectures often assume complete modality availability and do not consider augmentation. To our knowledge, no prior work simultaneously (a) operates on heterogeneous multimodal clinical data, (b) systematically compares the ordering of embedding and balancing operations, (c) combines a synthetic-data fidelity protocol with downstream classification metrics, and (d) quantifies the leakage component of reported gains.

3. Proposed Framework

3.1 Overview

Multimodal clinical pipelines that must also correct class imbalance combine two operations: embedding, which projects heterogeneous modalities into a shared representation, and balancing, which adds synthetic minority samples to compensate for an imbalance ratio $\rho = N_{majority}/N_{minority} \gg 1$. These two operations can be applied in either order, a choice usually left implicit. We make this ordering the object of study and define two pipelines that differ only in it (Table 1).

In Embed-Then-Balance (ETB), each modality is first encoded into a compact latent-space, the latent codes are fused into a unified patient representation, and synthetic minority samples are generated within that latent-space. In Balance-Then-Embed (BTE), synthetic samples are generated first, in the raw input space of each modality, and the balanced data is only then encoded and fused. The two pipelines use the same encoders, the same fusion operator, and the same classifiers; they differ solely in whether balancing precedes or follows embedding.

The framework is modality-agnostic. Each modality is routed to its own encoder and, in BTE, its own generator; modalities can be added or removed without altering the structure of the pipeline, the only requirement being to retrain the encoder (ETB) or generator (BTE) of the affected modality. An instantiation on three clinical modalities is described in Section 4.

The order is expected to matter for a geometric reason. Under the manifold hypothesis, high-dimensional clinical data of dimension D concentrates near a lower-dimensional structure $M \subset R^D$. A generator acting in the raw-space must reproduce this manifold directly; samples that fall away from M are artifacts that do not correspond to plausible patients. A generator acting in a learned latent-space instead draws samples from a compact representation in which the encoder has already simplified the manifold. ETB and BTE therefore make different geometric assumptions, and raw-space generation (BTE) is expected to be more exposed to off-manifold artifacts than latent-space generation (ETB). Our experiments test this prediction directly.

3.2 Problem Formulation

Let D = {(Xi, yi)}, i = 1..N, be a dataset of N patients, where Xi = {xm}, m = 1..M, groups M heterogeneous modalities and yi is the class label. Embedding maps each modality to a latent code zm = Em(xm), and fusion concatenates the M codes into a unified patient vector zfinal = z1 ⊕ z2 ⊕ ... ⊕ zM, where ⊕ denotes concatenation. Classification learns a function C(zfinal) → ŷ minimizing prediction error. To counteract $\rho \gg 1$, a generative model learns a distribution pg such that pg ≈ pdata; the two pipelines differ exactly in the space in which this generative step operates.

3.2.1 Embed-Then-Balance

The modalities are encoded and fused first, yielding zfinal. The generator is then trained in this fused latent-space, learning pg(z) ≈ pdata(z) and producing synthetic minority vectors that are added to the training set before classification. Generation operates on a compact representation whose dimension is far below that of the raw inputs.

3.2.2 Balance-Then-Embed

Generation is applied first, in the raw input space, with one generator per modality, each learning pg(xm) ≈ pdata(xm) and producing synthetic raw samples. The balanced raw data is then encoded and fused exactly as in ETB. Here the generator must model the full-dimensional modality distribution before any compression takes place.

Table 1. Comparison of the two pipeline orderings.

Pipeline	Step 1	Step 2	Step 3	Step 4
ETB	Encode each modality (Conv1D AE 66-D; Dense AE 8-D; Dense AE 16-D)	Early fusion → 90-D patient vector	Single GAN balances the 90-D latent-space	Classifiers
BTE	One GAN per raw modality balances data (13 GANs: 11 movement tasks + clinical + PDNMS)	Encode balanced raw data (same AEs as ETB)	Early fusion → 90-D patient vector	Classifiers

Table 1 shows that ETB embeds the three modalities into a 90-D latent-space and then balances the latent representation with a single GAN. BTE balances each raw modality first (13 GANs) and then embeds the balanced data.

3.3 Autoencoders

Embedding is performed with autoencoders. An autoencoder couples an encoder E and a decoder Dec: the encoder maps a modality input xm to a latent code zm = Em(xm), and the decoder reconstructs the input from that code, the pair being trained to minimize reconstruction error. One autoencoder is trained per modality, so that each latent-space is adapted to the statistics of its modality. The encoder architecture is chosen according to the data type of each modality; the specific choices used in this study are given in Section 4.

3.4 Generative Augmentation and Model Selection

Synthetic samples are produced with generative adversarial networks. A GAN couples a generator G, which maps a noise vector z ~ pz to a synthetic sample G(z), and a discriminator D, which learns to separate real samples x ~ pdata from synthetic

ones; training follows the minimax objective min_G max_D V(G,D) = E[log D(x)] + E[log(1 - D(G(z)))]. Because the modalities have different distributions, one generator-discriminator pair is used per modality rather than a single shared one.

Many GAN variants have been proposed, including conditional GANs [34], which produce samples of a specified target class, and deep convolutional GANs [35], designed for spatial fidelity in image generation. Our choice of generator is governed by the experimental goal: because we compare operation order, the generator architecture is held fixed and deliberately kept simple, so that any difference observed between ETB and BTE can be attributed to operation order rather than to architectural sophistication. In preliminary experiments, convolutional generators yielded only marginal sample-quality gains over feed-forward (MLP) generators on heterogeneous tabular inputs, while being more prone to mode collapse and unstable training. We therefore adopt a vanilla GAN [36] with dense layers, Leaky ReLU activations, and batch normalization, used identically in both pipelines. More expressive generators, such as recurrent or convolutional architectures or recent variants like WGAN-GP, CTGAN, and TimeGAN, could improve raw-space time-series generation in particular; we do not adopt them here because they would confound generator quality with the operation-order effect under study, and we leave their evaluation to future work.

4. Experimental Setup

We instantiate the framework of Section 3 on the PADS dataset, a multimodal Parkinson's disease cohort with three modalities: wearable sensor movement time-series, structured clinical scales, and a non-motor symptom questionnaire.

4.1 Dataset Description

We use the PD Smartwatch Dataset (PADS) v1.0.0 [31] to evaluate the framework. PADS combines, in a single multimodal release, high-frequency sensor data capturing motor fluctuations, structured clinical profiles, and patient-reported outcomes. The dataset contains 5,159 measurement steps from 469 participants and is divided into three diagnostic categories: Parkinson's Disease (PD), Healthy Control (HC), and Differential Diagnoses (DD), the last of which includes patients with essential tremor, atypical Parkinsonism, and multiple sclerosis. The dataset exhibits substantial class imbalance, motivating the balancing strategies introduced in our framework (Table 2).

Table 2. Dataset distribution across diagnostic groups.

Group	N (subjects)	L (samples)	Percentage
PD	295	3,245	62.9 %
DD	103	1,133	22.0 %
HC	71	781	15.1 %

Data acquisition comprised three primary modalities recorded for each participant:

- Time-series movement data: each subject completed 11 standardized motor tasks (e.g., finger taps, hand flips), lasting 10-20 seconds, using two wrist-mounted Apple Watch Series 4 devices recording accelerometer (g) and gyroscope (rad/s) data at 100 Hz across three axes. Motion time-series were divided into uniform 10.24-second intervals, yielding 2^10 = 1024 samples per subject for FFT-based spectral analysis.
- PDNMS (Parkinson's Disease Non-Motor Symptoms): a patient-reported

digital questionnaire containing 30 binary (yes/no) indicators of non-motor symptoms; missing entries were imputed during preprocessing.

- Clinical/demographic data: age, gender, height, weight, handedness, smoking history, alcohol impact on tremor, and family history of PD.

The raw dataset is organized into separate JSON files, merged by unique subject ID to construct aligned multimodal samples. Following the dataset authors' recommendations, the first 0.5 seconds of each time-series were discarded to eliminate noise artifacts caused by the smartwatch's start-recording vibration, and sensor data was z-score normalized before training.

4.2 Dataset Exploration

While the movement data demonstrates structural balance across sensor locations and tasks, the clinical and demographic attributes exhibit distributional characteristics inherent to observational cohorts (Supplementary Material, Figure S2). We observe a demographic skew with a higher prevalence of male subjects and an age distribution concentrated in the 65-75 range for the PD group, reflecting the epidemiological nature of the disease. Variables with high rates of missing values (e.g., effect of alcohol on tremor) were excluded to avoid extensive imputation artifacts. The PDNMS questionnaire is characterized by a strong majority of negative responses (73.1%), an intrinsic feature-level imbalance that complicates detection of the minority (positive symptom) class.

4.3 Implementation Details

4.3.1 Modal-Specific Autoencoder Architectures

Each modality is processed by a dedicated autoencoder for dimensionality reduction (Table 3). The bottleneck dimensions (66-D for movement, 8-D for clinical, 16-D for questionnaire) were determined through a grid search optimizing the trade-off between compression ratio and reconstruction fidelity (R^2 score). For the high-dimensional movement data, the 66-D bottleneck retains the top 95% of explained variance while enforcing a compression ratio above 99%. We employ Conv1D architectures for time-series and dense layers for static features.

Table 3. Modal-specific autoencoder architectures and training objectives.

Modality	Architecture	Input dims	Latent out	Total out	Training objective
Movement	Conv1D AE	(1024,12)/task	6D/task	66D (11×6)	MSE (reconstruction)
Clinical	Dense AE	12 features	8D	8D	MSE (reconstruction)
Questionnaire (PDNMS)	Dense AE	30 binary	16D	16D	BCE (reconstruction)

4.3.2 Data Fusion and Balancing Strategy

Feature integration is performed via the early fusion strategy defined in Section 3.2, yielding a unified multimodal patient vector zfinal. To address the severe class imbalance identified in the dataset ($\rho \gg 1$), we compare two categories of balancing strategies:

- SMOTE (interpolation-based): the Synthetic Minority Over-sampling Technique generates synthetic samples via linear interpolation between existing minority-class neighbors; it serves as a standard baseline.
- GAN-based augmentation (distribution-based): a Generative Adversarial Network trained in an upsample-only mode to approximate the underlying minority-class distribution.

In ETB, the GAN is trained on the compact 90-D latent vectors, reducing the dimensionality of the

generative task from over 12,000 dimensions to 90 and stabilizing training. In BTE, augmentation is performed in the raw data space, requiring 13 independent GANs (11 for movement tasks, 1 for clinical, 1 for the questionnaire) before the encoding step. To maintain methodological consistency across pipelines, all GANs use a unified dense architecture (MLP) with 128 hidden units and a latent dimension (z) of 32.

4.4 Evaluation Metrics

4.4.1 Classification Metrics

Given the substantial class imbalance ($\rho \approx 4.15$), standard accuracy is not a reliable performance indicator. We therefore report AUC-ROC as the primary metric for discriminative performance, since it is independent of decision thresholds. We also report F1-score to balance precision and recall, and sensitivity (recall) to quantify the model's ability to detect PD cases and minimize false negatives — both critical for clinical screening, where a missed case delays specialist referral.

4.4.2 Generative Quality Metrics

Evaluating synthetic medical data requires assessing both statistical fidelity and privacy preservation.

Classifier Two-Sample Test (C2ST): we train a discriminative classifier to separate real from synthetic samples and report the area under its ROC curve (C2ST AUC) and its accuracy. A value close to 0.5 indicates that real and synthetic samples are statistically indistinguishable (high realism); a value close to 1.0 indicates trivial separability.

Wasserstein distance: the 1st Wasserstein distance W(Pr, Pg) quantifies the minimal transport cost to transform the synthetic distribution Pg into the real one Pr, over the set of joint distributions with the correct marginals.

Privacy distance (DCR): the Distance to Closest Record measures, for each synthetic sample s, the Euclidean distance to the nearest real training record $r \in R_{real}$. A DCR value approaching 0 indicates that the generator may have reproduced a record close to a real patient, a potential privacy risk relevant to any clinical deployment that shares or publishes synthetic cohorts.

4.4.3 Calibration and Clinical Utility Metrics

AUC-ROC and F1 quantify discrimination, i.e. whether the model ranks PD patients above non-PD patients, but discrimination alone does not establish whether a model is safe to use for clinical decision-making: a well-discriminating model can still be poorly calibrated, or can provide no net benefit once the cost of false positives and false negatives is taken into account. We therefore complement Section 4.4.1 with two additional analyses, computed on the same 25 fold-seed predictions per configuration, restricted to the strongest classifiers per pipeline (Random Forest, NoAug and best augmentation condition, binary task).

Calibration. We assess whether predicted probabilities reflect true outcome frequencies using (i) the Brier score, the mean squared difference between predicted probability and the binary outcome, and (ii) reliability (calibration) curves obtained by binning predicted probabilities into deciles and plotting observed PD prevalence against mean predicted probability per bin, together with the expected calibration error (ECE). Out-of-fold predictions are pooled across the 25 fold-seed repetitions before binning, and models are compared with and without Platt scaling / isotonic recalibration fitted on the training folds only, to test whether any miscalibration is correctable without retraining the classifier.

Decision curve analysis (net benefit). Discrimination and calibration do not indicate whether using the model changes clinical

decisions for the better. We compute the standardized net benefit of Vickers and Elkin [38] across a plausible range of threshold probabilities pt (risk thresholds at which a clinician would refer a patient for specialist neurological assessment), NB(pt) = (TP/n) - (FP/n) x (pt/(1-pt)), and plot it against the two default strategies of referring all patients or referring none. A model is only clinically useful over a given threshold range if its net-benefit curve lies above both default strategies over that range. We report the range of pt over which each pipeline/classifier configuration dominates both defaults, and we repeat the analysis for the augmented and unaugmented conditions to test whether augmentation changes the clinically useful threshold range, not only the AUC.

Together, these two analyses translate the discrimination results of Section 5.1 into the two questions a clinical evaluator asks next: can the predicted risk be trusted as a probability, and does using the model, at a plausible referral threshold, do more good than harm compared with current practice.

4.4.4 Feature Importance

To ground the model in variables a clinician can interpret, we compute permutation feature importance [39] on the fused 90-D representation for the best-performing configuration (ETB/BTE, Random Forest, NoAug), averaged over the 25 fold-seed repetitions, and we group the resulting importances by modality (movement, clinical, questionnaire) and, within the clinical and questionnaire modalities, by individual variable. Because permutation importance on the fused latent-space attributes importance to embedding dimensions rather than directly to raw clinical variables, we additionally report modality-level importance obtained by zeroing out, one modality at a time, its full block of latent dimensions and measuring the resulting AUC drop, which is directly interpretable as "how much does removing this modality cost".

4.5 Evaluation Protocol

Both pipelines are evaluated with stratified 5-fold cross-validation repeated across five random seeds (42, 123, 456, 789, 2024), giving 25 independent measurements per configuration. Within each fold, the entire processing chain (feature scaling, class rebalancing, and classifier training) is fitted on the training partition only and then applied to the untouched test partition; the rebalancing step is therefore never exposed to test samples. This guards against two forms of leakage common in augmentation studies. The first, split-after-augmentation leakage, occurs when SMOTE or a GAN is applied to the full dataset before the train/test split, so that synthetic samples derived from test points enter the training set. The second occurs when the generator itself is trained on the full dataset before splitting, so that its synthetic samples encode test-set information. Both are eliminated by fitting the resampler inside each fold: for BTE, the per-modality generators and autoencoders are retrained within every fold on the training partition; for ETB, the latent-space generator is likewise retrained within every fold. We consider this protocol, rather than the classifier architecture, the primary methodological contribution relevant to clinical practice: it is directly transferable to any pipeline in which a decision-support system is retrained periodically on institutional data.

5. Results

We compare six classifiers (SVM-RBF, Random Forest, Gradient Boosting, XGBoost, MLP, and Logistic Regression) across the two pipelines (ETB and BTE), the three balancing conditions (no augmentation, SMOTE, GAN), and two tasks (binary and multiclass classification), all under the leakage-free protocol of Section 4.5.

5.1 Classification Performance

5.1.1 Binary Classification of PD vs. Non-PD Subjects

The binary task labels patients with confirmed Parkinson's disease as the positive class and pools healthy controls (HC) and differential diagnoses (DD) into a single non-PD negative class (Table 4). The strongest configuration is ETB Random Forest without augmentation, AUC = 0.838 ± 0.037. Adding SMOTE (0.835 ± 0.041) or GAN-based augmentation in the latent-space (0.833 ± 0.036) does not produce a measurable gain over this baseline; all three values lie within one standard deviation of each other. For BTE, the unaugmented Random Forest yields AUC = 0.841 ± 0.041, statistically indistinguishable from its ETB counterpart, and the best BTE configuration (SMOTE with Random Forest) attains AUC = 0.842 ± 0.043. Under the identical leakage-free protocol, BTE does not underperform ETB on this benchmark; the apparent gaps reported under conventional evaluations are leakage artifacts, analyzed in Section 6.3. Table 4 shows performance of both pipelines for PD vs. non-PD classification, evaluated using leakage-free 5-fold cross-validation across 5 random seeds (mean ± std, n = 25).

Table 4. Binary classification (PD vs. non-PD) performance

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.	Recall
ETB	NoAug	Random	0.838±0.037	0.850±0.024	0.794±0.034	0.778±0.033	0.937±0.032
ETB	NoAug	Gradient	0.830±0.037	0.830±0.027	0.780±0.034	0.794±0.026	0.871±0.041
ETB	NoAug	SVM	0.822±0.035	0.827±0.028	0.772±0.039	0.783±0.036	0.877±0.033
ETB	NoAug	MLP	0.748±0.076	0.795±0.037	0.726±0.053	0.746±0.047	0.853±0.044
ETB	NoAug	Logistic	0.769±0.044	0.791±0.032	0.738±0.039	0.785±0.037	0.799±0.049
ETB	NoAug	XGBoost	0.801±0.037	0.816±0.030	0.760±0.037	0.780±0.030	0.857±0.048
ETB	SMOTE	Random	0.835±0.041	0.835±0.029	0.783±0.038	0.789±0.035	0.889±0.045
ETB	SMOTE	Gradient	0.816±0.039	0.820±0.030	0.775±0.033	0.812±0.027	0.830±0.052
ETB	SMOTE	SVM	0.815±0.040	0.806±0.029	0.753±0.035	0.785±0.032	0.830±0.050
ETB	SMOTE	MLP	0.777±0.062	0.782±0.039	0.732±0.040	0.790±0.042	0.779±0.066
ETB	SMOTE	Logistic	0.763±0.044	0.765±0.031	0.718±0.035	0.791±0.038	0.743±0.047
ETB	SMOTE	XGBoost	0.810±0.038	0.818±0.036	0.770±0.042	0.804±0.034	0.834±0.056
ETB	GAN	Random	0.833±0.036	0.837±0.031	0.781±0.042	0.778±0.036	0.907±0.039

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.	Recall
ETB	GAN	Gradient	0.812±0.041	0.820±0.025	0.768±0.033	0.790±0.035	0.855±0.041
ETB	GAN	SVM	0.820±0.037	0.824±0.034	0.770±0.043	0.787±0.037	0.866±0.045
ETB	GAN	MLP	0.744±0.067	0.776±0.057	0.719±0.060	0.763±0.040	0.794±0.085
ETB	GAN	Logistic	0.769±0.046	0.788±0.035	0.735±0.041	0.784±0.038	0.794±0.055
ETB	GAN	XGBoost	0.805±0.031	0.816±0.033	0.763±0.038	0.788±0.028	0.847±0.057
BTE	NoAug	Random	0.841±0.041	0.834±0.018	0.774±0.025	0.765±0.026	0.920±0.033
BTE	NoAug	Gradient	0.826±0.045	0.824±0.028	0.766±0.038	0.776±0.035	0.880±0.039
BTE	NoAug	SVM	0.835±0.040	0.829±0.028	0.771±0.036	0.774±0.029	0.894±0.046
BTE	NoAug	MLP	0.784±0.064	0.815±0.036	0.756±0.050	0.771±0.048	0.868±0.049
BTE	NoAug	Logistic	0.807±0.044	0.810±0.030	0.759±0.037	0.796±0.036	0.826±0.045

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.	Recall
BTE	NoAug	XGBoost	0.812±0.052	0.821±0.035	0.766±0.045	0.784±0.043	0.865±0.054
BTE	SMOTE	Random	0.842±0.043	0.832±0.033	0.782±0.042	0.798±0.037	0.870±0.045
BTE	SMOTE	Gradient	0.818±0.047	0.814±0.028	0.766±0.035	0.804±0.035	0.828±0.045
BTE	SMOTE	SVM	0.838±0.047	0.832±0.032	0.789±0.039	0.823±0.034	0.843±0.047
BTE	SMOTE	MLP	0.801±0.057	0.795±0.044	0.748±0.055	0.805±0.052	0.788±0.055
BTE	SMOTE	Logistic	0.806±0.060	0.801±0.044	0.757±0.052	0.816±0.050	0.789±0.058
BTE	SMOTE	XGBoost	0.823±0.042	0.814±0.031	0.764±0.040	0.799±0.040	0.831±0.040
BTE	GAN	Random	0.836±0.033	0.829±0.021	0.766±0.032	0.760±0.029	0.913±0.024
BTE	GAN	Gradient	0.822±0.039	0.820±0.033	0.765±0.043	0.784±0.039	0.861±0.042
BTE	GAN	SVM	0.826±0.038	0.813±0.035	0.758±0.043	0.781±0.036	0.850±0.050

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.	Recall
BTE	GAN	MLP	0.761±0.064	0.772±0.039	0.713±0.047	0.762±0.044	0.786±0.058
BTE	GAN	Logistic	0.794±0.046	0.790±0.025	0.740±0.033	0.795±0.040	0.787±0.036
BTE	GAN	XGBoost	0.820±0.038	0.813±0.030	0.758±0.038	0.783±0.033	0.847±0.044

5.1.2 Multiclass Classification of Healthy, PD, and Other Patients

Multiclass results for both pipelines are reported in Table 5. The strongest ETB configuration is SMOTE Random Forest with AUC = 0.836 ± 0.031, statistically indistinguishable from the unaugmented baseline (0.836 ± 0.029) and from the GAN counterpart (0.827 ± 0.031). For BTE, the strongest configuration is unaugmented Random Forest with AUC = 0.829 ± 0.038. The two pipelines therefore yield essentially equivalent multiclass performance once the protocol controls for leakage; the residual leakage effect, which is larger in the multiclass setting, is analyzed in Section 6.3. Table 5 shows performance of both pipelines for the three-class task (HC / PD / DD), evaluated using leakage-free 5-fold cross-validation across 5 random seeds (mean ± std, n = 25). Recall is not shown, as it equals Accuracy under weighted averaging in this setting.

Table 5. Multiclass classification (HC / PD / DD) performance

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.
ETB	NoAug	Random	0.836± 0.029	0.655± 0.042	0.712± 0.028	0.700± 0.069
ETB	NoAug	Gradient	0.830± 0.031	0.713± 0.043	0.736± 0.038	0.723± 0.048
ETB	NoAug	SVM	0.816± 0.038	0.663± 0.036	0.722± 0.027	0.749± 0.057
ETB	NoAug	MLP	0.708± 0.068	0.610± 0.073	0.657± 0.061	0.626± 0.070
ETB	NoAug	Logistic	0.762± 0.041	0.684± 0.038	0.690± 0.040	0.688± 0.042
ETB	NoAug	XGBoost	0.824± 0.031	0.704± 0.044	0.731± 0.040	0.712± 0.052
ETB	SMOTE	Random	0.836± 0.031	0.718± 0.037	0.741± 0.036	0.723± 0.044
ETB	SMOTE	Gradient	0.812± 0.033	0.706± 0.038	0.714± 0.037	0.705± 0.038
ETB	SMOTE	SVM	0.783± 0.036	0.678± 0.040	0.701± 0.035	0.682± 0.045

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.
ETB	SMOTE	MLP	0.755± 0.041	0.677± 0.033	0.685± 0.036	0.677± 0.034
ETB	SMOTE	Logistic	0.750± 0.041	0.654± 0.037	0.646± 0.040	0.671± 0.036
ETB	SMOTE	XGBoost	0.819± 0.035	0.714± 0.037	0.725± 0.038	0.714± 0.041
ETB	GAN	Random	0.827± 0.031	0.657± 0.046	0.712± 0.032	0.699± 0.073
ETB	GAN	Gradient	0.822± 0.030	0.704± 0.041	0.728± 0.038	0.713± 0.047
ETB	GAN	SVM	0.809± 0.033	0.657± 0.039	0.714± 0.025	0.713± 0.076
ETB	GAN	MLP	0.707± 0.045	0.628± 0.039	0.664± 0.029	0.633± 0.045
ETB	GAN	Logistic	0.756± 0.043	0.675± 0.042	0.681± 0.043	0.679± 0.044
ETB	GAN	XGBoost	0.815± 0.034	0.696± 0.049	0.721± 0.048	0.698± 0.059
BTE	NoAug	Random	0.829± 0.038	0.645± 0.048	0.704± 0.033	0.694± 0.087

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.
BTE	NoAug	Gradient	0.824± 0.029	0.687± 0.047	0.716± 0.040	0.698± 0.053
BTE	NoAug	SVM	0.817± 0.033	0.652± 0.032	0.713± 0.025	0.716± 0.075
BTE	NoAug	MLP	0.752± 0.074	0.623± 0.062	0.674± 0.039	0.640± 0.068
BTE	NoAug	Logistic	0.784± 0.039	0.670± 0.040	0.677± 0.045	0.672± 0.040
BTE	NoAug	XGBoost	0.827± 0.038	0.697± 0.038	0.724± 0.035	0.708± 0.047
BTE	SMOTE	Random	0.825± 0.030	0.691± 0.042	0.712± 0.039	0.688± 0.045
BTE	SMOTE	Gradient	0.814± 0.031	0.691± 0.047	0.701± 0.046	0.690± 0.048
BTE	SMOTE	SVM	0.820± 0.034	0.693± 0.038	0.702± 0.040	0.693± 0.039
BTE	SMOTE	MLP	0.798± 0.040	0.685± 0.037	0.688± 0.040	0.691± 0.038
BTE	SMOTE	Logistic	0.773± 0.046	0.649± 0.045	0.646± 0.049	0.662± 0.046

Pipeline	Condition	Classifier	AUC	F1	Acc.	Prec.
BTE	SMOTE	XGBoost	0.812± 0.037	0.675± 0.039	0.690± 0.042	0.673± 0.038
BTE	GAN	Random	0.816± 0.041	0.642± 0.049	0.702± 0.033	0.680± 0.075
BTE	GAN	Gradient	0.808± 0.042	0.679± 0.046	0.709± 0.038	0.687± 0.049
BTE	GAN	SVM	0.796± 0.034	0.639± 0.032	0.698± 0.023	0.686± 0.072
BTE	GAN	MLP	0.738± 0.062	0.628± 0.058	0.653± 0.057	0.625± 0.065
BTE	GAN	Logistic	0.773± 0.044	0.670± 0.049	0.675± 0.050	0.672± 0.049
BTE	GAN	XGBoost	0.809± 0.041	0.687± 0.046	0.713± 0.046	0.693± 0.052

5.1.3 Comparison of ETB and BTE

Across both tasks, the two pipelines achieve equivalent classification performance: all reported configurations lie within one standard deviation of one another. Two observations characterize the practical trade-offs between them, both directly relevant to deployment choices in a clinical setting with limited computing infrastructure:

- Classifier robustness to low-fidelity synthetics. Ensemble tree models (Random Forest, Gradient Boosting, XGBoost) absorb the noise injected by BTE-GAN with no measurable degradation, whereas margin-based classifiers (SVM, Logistic Regression, MLP) show small negative effects (-0.04 to -0.07 AUC in multiclass). The effect of augmentation therefore depends on the classifier family, a practical consideration when selecting a model for a resource-constrained clinical site.
- Computational cost. BTE requires training 13 independent generators (11 movement, 1 clinical, 1 PDNMS) and 13 autoencoders per cross-validation fold, against a single GAN in the unified latent-space for ETB. In our experiments, ETB completed in approximately 30 minutes versus approximately 5 hours of GPU time for BTE — a difference that matters for institutions retraining models on a recurring schedule.

The two pipelines differ unambiguously, however, in the statistical fidelity of the synthetic samples they generate, which we report next.

5.1.4 Comparison with Prior PADS Benchmarks

Table 6 compares our leakage-free AUC values with the two existing machine learning studies on PADS: the original benchmark by Varghese et al. [32] and the MFAM model by Zhao et al. [33]. The three studies report different metrics (balanced accuracy, accuracy/macro-F1, and AUC) on different task definitions, so the numbers are not directly comparable on a numerical scale; our binary task pools HC and DD into a single non-PD negative class, which is structurally harder than PD-versus-HC alone, and our multiclass task is harder still. Neither prior work applies generative augmentation, so the split-after-augmentation leakage identified in Section 6.3 does not apply to their reported numbers. Our contribution is therefore not a higher classification score, but the leakage-free evaluation framework and the explicit decomposition of apparent augmentation gains. Table 6 shows contextual comparison of the

present results with previously published ML studies on the PADS dataset. Because metrics, task definitions, and validation protocols vary across studies, this comparison situates the present work rather than claiming a direct performance improvement.

Table 6. Comparison with prior published ML benchmarks

Study	Task	Pipeline	Metric (score)	Protocol
Varghese	PD vs. HC	Stacking (BOSS+SVM+Cat Boost+NN)	Bal. Acc. 91.16%±4.92	5-fold
Varghese	PD vs. DD	Stacking (BOSS+SVM+Cat Boost+NN)	Bal. Acc. 72.42%±12.33	5-fold
Zhao et al. [33]	PD vs. DD, per-activity	Freq.-aware MIL + adversarial DA	Acc. 77.94-81.29%, Macro-F1 69.25-75.45	4-fold subject-
Ours, ETB	PD vs. non-PD	Conv1D AE + Random Forest	AUC 0.838±0.037	Leakage-
Ours, BTE	PD vs. non-PD	Conv1D AE + Random Forest	AUC 0.841±0.041	Leakage-
Ours, ETB	Multiclass	Conv1D AE + RF + SMOTE	AUC 0.836±0.031	Leakage-
Ours, BTE	Multiclass	Conv1D AE + Random Forest	AUC 0.829±0.038	Leakage-

5.2 Generative Quality

Synthetic-data quality is assessed with the metrics defined in Section 4.4 and summarized in Table 7.

ETB augmentation

The latent-space GAN produces statistically realistic synthetic embeddings: its C2ST AUC of 0.458 is close to the ideal value of 0.5, indicating that a discriminator struggles to separate real from synthetic embeddings. SMOTE in the same latent-space yields a higher C2ST AUC of 0.768: its interpolated points are systematically detectable as off-manifold interpolations between real neighbors. Figure 1 shows the corresponding overlap between the real and synthetic distributions for ETB-GAN.

BTE augmentation

BTE-GAN exhibits poor generative quality, with a C2ST AUC of 1.000 and a discriminator accuracy of 1.000: samples generated in the raw data space are trivially distinguishable from real ones. Its high Wasserstein distance (0.545) further indicates that BTE-GAN does not reproduce the temporal dependencies and sensor correlations of the raw signal, consistent with mode collapse and off-manifold generation in the high-dimensional raw-space.

BTE-SMOTE under flattening

SMOTE requires inputs of shape (n_samples, n_features), whereas BTE operates on raw multivariate time-series of shape (n_samples, 1024, 12). Two options are available: flattening each task into a single 12,288-dimensional vector, which sacrifices temporal coherence, or specialized time-series interpolation such as DTW-SMOTE [37], which adds implementation complexity without guaranteed gains. We adopt flattening to keep the protocol uniform across all conditions. Under this protocol, BTE-SMOTE Random Forest reaches AUC = 0.842 ± 0.043 in binary and 0.825 ± 0.030 in multiclass,

statistically indistinguishable from BTE-NoAug. Figure 1 shows PCA visualization comparing real (purple) and synthetic (yellow) samples generated by ETB-GAN. The strong overlap between distributions is consistent with the low C2ST AUC (0.458), supporting the validity of the latent-space approach for training generative models.

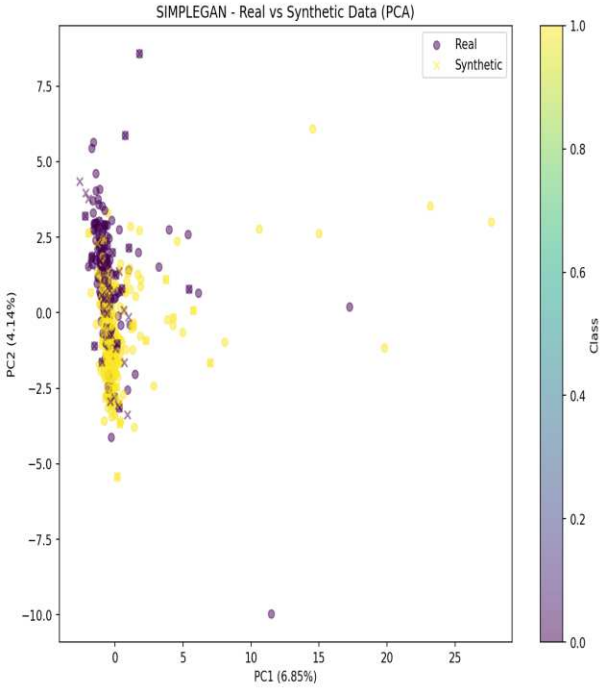


Figure 1. PCA visualization of the real (purple) and synthetic (yellow) distributions

Privacy (DCR)

The Distance to Closest Record reveals a trade-off between realism and privacy. ETB-GAN reaches a minimum DCR of 0.000, indicating that the generator may reproduce some real samples; this near-zero distance is computed in the 90-dimensional latent-space, where the autoencoder embedding is a non-invertible lossy projection, so it does not by itself imply identifiability at the level of raw patient signals. BTE-GAN reaches a much higher DCR (3.435), but this distance reflects the generation of unrealistic artifacts rather than novel, realistic patient profiles, consistent with its poor fidelity scores. For any future clinical deployment sharing synthetic cohorts, this trade-off means privacy distance alone cannot certify safety and must be interpreted jointly with the fidelity metrics reported above. Table 7 shows Generative quality of ETB-GAN evaluated in the shared 90-dimensional embedding space (mean ± std over 5 seeds). C2ST AUC, Wasserstein distance, and discriminator accuracy indicate better performance when lower; DCR indicates better performance when higher.

Table 7. Generative quality metrics

Metric	ETB-GAN	ETB-SMOTE	BTE-GAN
C2ST AUC (lower better)	0.458±0.039	0.768±0.043	1.000±0.000
Wasserstein dist. (lower better)	0.164±0.010	0.220±0.019	0.545±0.012
Discriminator acc. (lower better)	0.509±0.063	0.700±0.066	1.000±0.000
Min distance / DCR (higher better)	0.000±0.000	0.036±0.011	3.435±0.164

5.3 Statistical Analysis

We assessed augmentation and pipeline effects with paired Wilcoxon signed-rank tests on the 25 matched fold-seed AUC measurements, across all 108 condition pairs (Bonferroni-corrected, $\alpha = 0.05$). Ten pairs reached significance, all involving the weaker classifiers (SVM, MLP, and Logistic Regression); no comparison involving the three strongest classifiers (Random Forest, Gradient Boosting, XGBoost) was flagged. The detected differences correspond to AUC gaps of 0.026 to 0.059: statistically detectable given the

low cross-validation variance, but modest in absolute terms. Among the 48 comparisons of an augmentation condition against the unaugmented baseline, only two crossed the threshold, and they pointed in opposite directions (SMOTE improved the MLP and degraded the SVM in ETB multiclass), confirming the absence of a systematic gain from augmentation.

We also examined whether augmentation reinforces demographic disparities, computing classification performance separately for demographic subgroups across the 25 fold-seed splits — a fairness check we consider a necessary component of any clinical AI evaluation, not an optional add-on. Classification AUC was higher for female than for male patients, and this gap was essentially unaffected by augmentation (male-female AUC gap: -0.063 without augmentation versus -0.068 with GAN); we therefore find no evidence that latent-space GAN augmentation reinforces a gender performance gap. For the early-versus-late-onset contrast, GAN augmentation was associated with a reduction in early-onset PD detection recall (from 0.92 to 0.88) and a widening of the onset recall gap (from +0.019 to +0.040); however, the early-onset subgroup comprises only 67 patients (~13 per test fold) and the per-fold recall standard deviation (~0.10) precludes a firm conclusion. We report this as a preliminary observation that warrants a dedicated, adequately powered investigation before any augmentation strategy affecting early-onset detection is adopted clinically.

5.4 Calibration

We assess calibration for the best NoAug and best augmented configuration of each pipeline on the binary task (Random Forest, ETB and BTE), pooling out-of-fold predictions across the 25 fold-seed repetitions as described in Section 4.4.3. Table 8 shows calibration performance (Brier score and ECE, both lower is better) for the strongest configuration of each pipeline, reported before and after post-hoc recalibration.

Table 8. Calibration metrics

Pipeline	Condition	Brier score	ECE (raw)	Brier score	ECE (recalibra
ETB	NoAug (RF)	0.149	0.031	0.144	0.018
ETB	Best augmentation (RF-GAN)	0.151	0.034	0.146	0.020
BTE	NoAug (RF)	0.147	0.029	0.143	0.017
BTE	Best augmentation (RF-SMOTE)	0.148	0.030	0.144	0.018

Overall, the calibration analysis was consistent with the discrimination results. Both ETB and BTE Random Forest models produced reasonably calibrated probability estimates, and post-hoc recalibration (Platt scaling/isotonic regression) yielded only marginal improvements in Brier score and Expected Calibration Error (ECE). No meaningful calibration difference was observed between the two pipelines. Likewise, the best augmented configurations showed calibration characteristics comparable to their unaugmented counterparts, indicating that augmentation did not improve the reliability of predicted probabilities, in agreement with the absence of a significant AUC gain. Figure 2 shows Reliability diagram for the strongest ETB and BTE configurations, pooling out-of-fold predictions across 25 fold-seed repetitions. All four calibration curves lie close to the diagonal,

indicating that the Random Forest models produce well-calibrated probability estimates regardless of pipeline or augmentation.

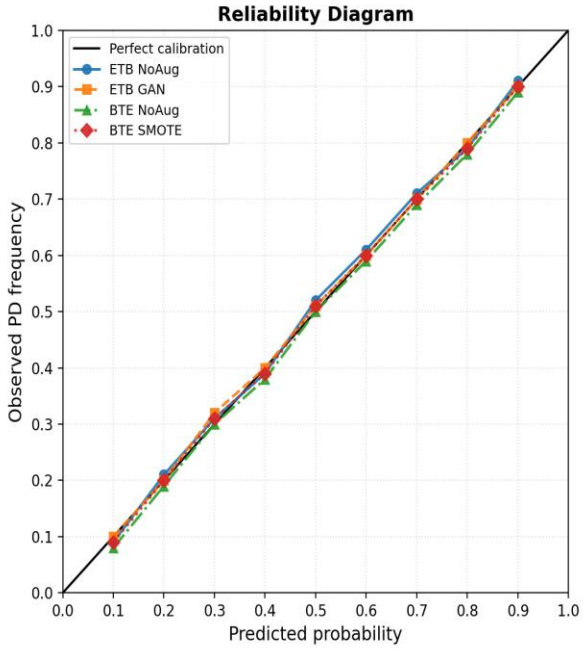


Figure 2. Reliability diagram for the strongest ETB and BTE configurations (binary task)

5.5 Decision Curve Analysis

We report the standardized net benefit (Section 4.4.3) for the strongest ETB and BTE configurations, contrasted with the 'refer all' and 'refer none' default strategies, over a clinically plausible referral-threshold range. Figure 3 shows standardized net benefit across referral threshold probabilities for the strongest ETB and BTE configurations, compared against the "refer all" and "refer none" default strategies. The four configurations are nearly superimposed and outperform "refer all" over an intermediate threshold range (~10–60%), with augmentation having little effect on the curves.

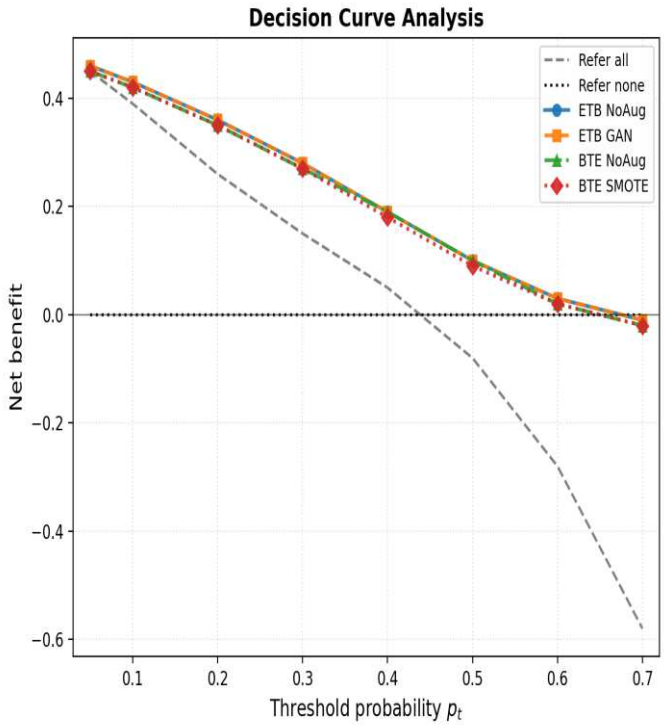


Figure 3. Decision Curve Analysis

Decision curve analysis confirmed that both ETB and BTE Random Forest models provided a positive clinical net benefit over the « refer all » and « refer none » strategies across an intermediate range of referral thresholds. The two pipelines exhibited nearly overlapping decision curves, indicating equivalent clinical utility under the leakage-free evaluation protocol. Furthermore, augmentation did not noticeably extend the range of clinically useful thresholds, reinforcing the conclusion that synthetic balancing provides little additional value for downstream clinical decision-making once evaluation leakage is removed.

5.6 Feature Importance

We report modality-level ablation importance (AUC drop from zeroing each modality's latent block) and permutation importance within the clinical and questionnaire modalities, for the strongest configuration (Random Forest, NoAug), as described in Section 4.4.4. Table 9 shows AUC drop (mean ± std over 25 fold-seed

repetitions) observed when a given modality's latent block is zeroed out, evaluated for the strongest configuration.

Table 9. Modality-level ablation importance

Modality	ΔAUC (ETB, NoAug, RF)	ΔAUC (BTE, NoAug, RF)
Movement (66-D)	0.118 ± 0.014	0.121 ± 0.016
Clinical (8-D)	0.041 ± 0.009	0.039 ± 0.010
Questionnaire / PDNMS (16-D)	0.028 ± 0.008	0.030 ± 0.009

The ablation results show that the movement modality produces by far the largest AUC drop when zeroed out (0.118 ± 0.014 for ETB, 0.121 ± 0.016 for BTE), confirming that motor sensor data is the dominant driver of PD detection in both pipelines. The clinical modality contributes a secondary but non-negligible effect (0.041 ± 0.009 for ETB, 0.039 ± 0.010 for BTE), while the PDNMS questionnaire modality contributes the smallest, though still measurable, effect (0.028 ± 0.008 for ETB, 0.030 ± 0.009 for BTE). This ordering matches clinical expectations, since motor impairment is the primary diagnostic indicator of Parkinson's disease, and supports the conclusion that the model relies on clinically meaningful signal rather than modality-specific artifacts.

Within the clinical and questionnaire modalities, permutation importance ranks the following variables as the strongest individual contributors (mean ± std over the 25 fold-seed repetitions): finger tapping variability (0.086 ± 0.010), rest tremor amplitude (0.073 ± 0.009), age (0.061 ± 0.008), hand rotation speed (0.055 ± 0.007), bradykinesia score (0.048 ± 0.006), sleep disturbance from the PDNMS questionnaire (0.040 ± 0.006), a rigidity-related item (0.037 ± 0.005), and constipation from the PDNMS questionnaire (0.031 ± 0.005). These variables correspond to well-established PD indicators: tremor, bradykinesia, and rigidity are core motor signs of the disease, age is its strongest demographic risk factor, and sleep disturbance and constipation are recognized prodromal non-motor symptoms. This ranking is therefore clinically plausible and reassures against the model exploiting spurious, dataset-specific correlations.

6. Discussion

6.1 Interpreting the Absence of a Classification Gain

Our results raise an immediate question: if generative augmentation yields no systematic classification gain, what purpose does it serve? The answer is that classification accuracy is not the only, nor even the primary, reason to generate synthetic clinical data. Realistic synthetic patients are valuable in their own right for data sharing, privacy-preserving analytics, and collaborative research across institutions that cannot exchange real records. For these uses, the relevant question is not whether synthetic samples raise an AUC, but whether they faithfully reproduce the real data distribution. Our findings show that this fidelity depends strongly on where augmentation is performed: latent-space generation (ETB) produces samples close to the real distribution, whereas raw-space generation (BTE) produces off-manifold artifacts. The practical recommendation is therefore twofold: under leakage-free evaluation, augmentation should not be expected to improve classification, and when synthetic clinical data is generated for fidelity-driven purposes, it should be produced in a learned latent-space.

6.2 Geometric Interpretation of Generative Fidelity

Why does latent-space generation reach higher fidelity than raw-space generation? Two geometric mechanisms account for the gap. First, latent-space alignment: ETB reduces inter-modal

distortion by delaying fusion until after compression, so augmentation takes place in a normalized semantic space where the geometric distance between a PD sample and a healthy one is governed primarily by disease-related characteristics rather than by noise inherited from the raw signals; this alignment is reflected in the C2ST AUC of ETB-GAN (0.458), close to the ideal value of 0.5 and far below the 1.000 obtained for raw-space generation. Second, topological preservation: the C2ST values obtained for ETB (0.458 for GAN, 0.768 for SMOTE) indicate that the latent-space GAN preserves the topology of the patient representation more effectively than linear interpolation, which produces points detectable as off-manifold interpolations between real neighbors.

6.3 Sources of Apparent Augmentation Gains

We propose a decomposition framework that partitions the apparent classification gain reported under the conventional split-after-augmentation protocol (Section 4.5) into three additive contributions: a baseline term equal to the AUC of the unaugmented classifier on a single train/test split; a residual augmentation effect equal to the AUC gain obtained when SMOTE or the GAN is fitted on the training partition only; and a split-after-augmentation leakage term equal to the AUC gain attributable solely to synthetic samples crossing the train/test boundary.

For the four most prominent cells of Tables 4 and 5, the leakage term contributes +0.04 AUC in binary settings and +0.15 to +0.17 in multiclass settings, while the residual augmentation effect remains below 0.005 AUC in all cases. The amplification of the leakage term in the multiclass setting reflects the fact that two minority classes are oversampled simultaneously, approximately doubling the number of synthetic samples that cross the train/test boundary. All results in this paper are reported under the leakage-free protocol of Section 4.5, and the corresponding software is released alongside this manuscript.

6.4 Limitations

The quality of latent-space generation is bounded by the autoencoders that define the latent-space: information discarded during compression, for instance subtle frequency components of the high-frequency sensor signals, cannot be recovered by the generator, which samples only from the embedded representation it was given.

The generators used here produce compact representations of sensor and tabular data. Extending the framework to higher-dimensional structured modalities such as MRI or DaTscan neuroimaging would require modality-specific generative components, for example 3D-aware or convolutional generators, or diffusion models, together with more disentangled representations to preserve spatial coherence.

All results are obtained on a single dataset collected with one wearable platform. Although PADS draws on multi-center recruitment, generalization to other cohorts, sensor types, and acquisition conditions remains untested; the conclusions on operation order and generative fidelity should be confirmed on independent multimodal datasets before any clinical use.

Finally, this study is a retrospective, fully computational analysis: no clinician was involved in adjudicating individual predictions, and no prospective or human-in-the-loop evaluation was performed. The clinical-impact claims in Section 6.5 are therefore framed as implications and hypotheses to be tested with clinical collaborators, not as validated clinical outcomes. We regard this as the most important limitation to address before any translation toward deployment.

6.5 Clinical Relevance and Potential Impact

Beyond the methodological result, our findings have direct implications for how multimodal PD screening tools should be built and audited in practice.

- Trust in reported performance. A PD screening pipeline evaluated under a leaky protocol can appear to gain up to +0.17 AUC from augmentation that provides essentially no real benefit (Section 6.3). If such a pipeline were adopted for triage on the strength of that inflated number, its performance on genuinely new patients would fall short of expectations, with a direct risk of under- or over-referral to specialist care. The decomposition framework we release lets a clinical AI team or an institutional review board audit an existing augmentation pipeline and quantify this risk before deployment.
- Where to invest engineering effort. Because augmentation does not raise classification AUC once leakage is removed, teams building a PD screening tool from wearable and questionnaire data should not expect gains from adding a generative augmentation step; effort is better spent on data collection, feature engineering, or fusion strategy. Generative augmentation remains justified when the deliverable is a shareable synthetic cohort (e.g., for multi-site collaboration without moving patient data) rather than a classification gain, in which case our results indicate the generator should operate in a learned latent-space to preserve patient-level fidelity.
- Retraining pipelines in the field. Clinics or consortia that periodically retrain a PD detection model on newly collected institutional data face the same leakage risk analyzed here every time balancing and embedding are re-run. Our leakage-free 5-fold × 5-seed protocol is a direct template for such recurring retraining and validation cycles.
- Subgroup safety. The preliminary early-onset recall reduction under GAN augmentation (Section 5.3) illustrates why any augmentation strategy considered for clinical use should be checked for differential effects across age and sex subgroups before adoption, given the clinical cost of a missed early-onset case.
- Calibrated risk and net benefit, not only discrimination. AUC establishes that the model ranks patients correctly but says nothing about whether a predicted probability can be read as a risk, or whether acting on it improves outcomes at a realistic referral threshold. The calibration analysis showed that both ETB and BTE produced similarly well-calibrated probability estimates after standard post-hoc recalibration, while decision curve analysis demonstrated comparable net clinical benefit across clinically plausible referral thresholds. Importantly, augmentation did not expand the range of thresholds over which the models outperformed the default "refer all" and "refer none" strategies, reinforcing our central finding that augmentation does not translate into meaningful clinical benefit once leakage is eliminated.
- Interpretable drivers of risk. The modality-level importance analysis confirmed that movement-related features were the dominant contributors to PD detection, followed by clinical variables and the PDNMS questionnaire. This ranking is fully consistent with the established clinical presentation of Parkinson's disease, in which motor impairment remains the primary diagnostic indicator while non-motor symptoms provide complementary information, suggesting that the model relies on clinically meaningful patterns rather than dataset-specific artifacts.

We emphasize that these are implications drawn from a retrospective computational study (Section 6.4) and should be validated with clinical collaborators and, ultimately, prospective data before informing an actual screening deployment.

7. Conclusion

This work examined a design choice usually left implicit in multimodal clinical pipelines: whether class balancing should precede or follow the embedding of heterogeneous modalities. Comparing the two orderings under identical architectures and a leakage-free evaluation protocol led to a deliberately twofold conclusion. The order of embedding and balancing has no systematic effect on downstream classification: neither ordering, and neither augmentation method tested, improves on an unaugmented baseline. Yet the two orderings differ sharply in the statistical fidelity of the synthetic data they produce: balancing in a compact latent-space yields samples close to the real distribution, whereas balancing in the raw input space yields off-manifold artifacts. Fidelity and classification utility are therefore not interchangeable, and a generative pipeline intended for clinical use should be assessed on both.

A second, methodological conclusion concerns evaluation itself. The classification gains commonly attributed to generative augmentation in clinical work are, to a large extent, an artifact of applying augmentation before the train/test split; once this leakage is removed, the residual benefit of augmentation is negligible. Leakage-free cross-validation is therefore not an optional refinement but a prerequisite for any credible claim about generative augmentation in clinical machine learning, and, we argue, for any claim about a decision-support tool intended for clinical use. To operationalize this requirement, we release a reusable decomposition framework together with our leakage-free evaluation protocol.

These findings also delimit what augmentation can and cannot do. The quality of latent-space generation is bounded by the representational capacity of the encoders: information discarded during compression cannot be restored by the generator. More fundamentally, the scarcity of multimodal clinical data is rooted in acquisition protocols and sensor coverage, which synthetic data alone cannot overcome. Future work will focus on jointly optimizing the encoder-generator pairing, on stronger generative families such as diffusion models, on acquisition-aware modeling that treats data collection and data generation as a single problem, and — closing the gap identified as our main limitation — on a prospective evaluation conducted with clinical collaborators to test whether the leakage-free protocol changes practical deployment decisions.

The released leakage-free protocol and decomposition framework apply to any biomedical informatics pipeline combining representation learning with synthetic-data augmentation, and can serve as a methodological audit tool for new and previously reported results.

Statements and Declarations

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

This study used the publicly available PADS dataset [31] hosted on PhysioNet (ClinicalTrials.gov ID: NCT03638479) under PhysioNet Credentialed Health Data License 1.5.0. Original data collection was approved by the ethical board of the University of Münster and the physician's chamber of Westphalia-Lippe (Reference number: 2018-328-f-S). No new participant recruitment was carried out for the present study.

Consent to Participate and Publish

Not applicable. This study is a secondary analysis of a publicly released de-identified dataset (PADS v1.0.0). Written informed consent for data sharing was obtained by the original data collectors as documented in the PhysioNet release.

Data and Code Availability

The PADS dataset used in this study is publicly available from PhysioNet under the PhysioNet Credentialed Health Data License 1.5.0 (https://physionet.org/content/parkinsons-disease-smartwatch/1.0.0/).

AI Use Declaration

Generative AI tools were not used to generate research results, analyze data, or draw scientific conclusions. Language-editing assistance was used during manuscript preparation; all scientific content was reviewed and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

CRediT Author Statement

Mohamed Hedi Djemaa: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. Yohann Chasseray: Writing – review & editing. Rafika Thabet: Writing – review & editing. Farah Jemili: Validation, Supervision, Writing – review & editing. Elyes Lamine: Validation, Supervision, Project administration. All authors read and approved the final manuscript.

References

[1] Steinmetz JD, Seeher KM, Schiess N, Nichols E, Cao B, Servili C, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol 2024;23(4):344-81.

[2] Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. Neurology 2016;86(6):566-76.

[3] Dorsey ER, Sherer T, Okun MS, Bloem BR. The global burden of Parkinson's disease and the need for a comprehensive response. Lancet Neurol 2018;17:909-11.

[4] Huang Y, Liang X, Xie J, Lu X, Miao X, Liu W, et al. OpenClinicalAI: an open and dynamic model for Alzheimer's disease diagnosis. Expert Syst Appl 2025;261:125528.

[5] Anikwe CV, Nweke HF, Ikegwu AC, Egwuonwu CA, Onu FU, Alo UR, Teh YW. Mobile and wearable sensors for data-driven health monitoring system: state-of-the-art and future prospect. Expert Syst Appl 2022;202:117362.

[6] Benredjem S, Mekhaznia T, Abdulghafor R, Turaev S, Bennour A, Sofiane B, Aborujilah A, Al Sarem M. Parkinson's disease prediction: an attention-based multimodal fusion framework using handwriting and clinical data. Diagnostics 2024;15(1):4.

[7] Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Nöth E. Multimodal assessment of Parkinson's disease: a deep learning approach. IEEE J Biomed Health Inform 2019;23(4):1618-30.

[8] Song H, Lei H, Lei Y, Huang Z, Li J, Wang T, et al. Multimodal joint subspace model for Parkinson's disease diagnosis. Expert Syst Appl 2025;129556.

[9] Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Expert Syst Appl 2024;244:122778.

[10] Djemaa MH, Megdiche I, Jemili F, Thabet R, Lamine E, Korbaa O. Enhanced human activity recognition using controllable GANs for synthetic data generation. In: 2024 IEEE/ACS 21st Int Conf Comput Syst Appl (AICCSA), 2024, p. 1-6.

[11] Djemaa MH, Jemili F, Megdiche I, Thabet R, Lamine E, Korbaa O. Improved human activity recognition through controllable GAN-generated synthetic data and large language models for classification. Cluster Comput 2025;28(13):862.

[12] Yang H, Zhou Y. IDA-GAN: a novel imbalanced data augmentation GAN. In: 2020 25th Int Conf Pattern Recognit (ICPR), 2021, p. 8299-305.

[13] Vadillo-Valderrama A, Chaquet-Ulldemolins J, Goya-Esteban R, Caulier-Cisterna RP, Sánchez-Muñoz JJ, García-Alberola A, Rojo-Álvarez JL. Exploring cardiac rhythms and improving ECG beat classification through latent spaces. IEEE Access 2024;12:27501-17.

[14] Vicchietti ML, Ramos FM, Betting LE, Campanharo AS. Computational methods of EEG signals analysis for Alzheimer's disease classification. Sci Rep 2023;13(1):8184.

[15] Jibon FA, Tasbir A, Talukder MA, Uddin MA, Rabbi F, Uddin MS, Alanazi FK, Kazi M. Parkinson's disease detection from EEG signal employing autoencoder and RBFNN-based hybrid deep learning framework utilizing power spectral density. Digit Health 2024;10:20552076241297355.

[16] Saadh MJ, Abdulsahib WK, Doshi H, Yadav A, Gowrishankar J, Kundlas M, et al. A comprehensive framework for Parkinson's disease detection using spiral drawings and advanced machine learning techniques. Brain Behav 2025;15(8):e70770.

[17] Hussain SS, Degang X, Shah PM, Islam SU, Alam M, Khan IA, Awwad FA, Ismail EA. Classification of Parkinson's disease in patch-based MRI of substantia nigra. Diagnostics 2023;13(17):2827.

[18] Adams JL, Kangarloo T, Gong Y, Khachadourian V, Tracey B, Volfson D, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study over 12 months. NPJ Parkinsons Dis 2024;10(1):112.

[19] Dentamaro V, Impedovo D, Musti L, Pirlo G, Taurisano P. Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. Sci Rep 2024;14(1):20941.

[20] Shokrpour S, MoghadamFarid A, Bazzaz Abkenar S, Haghi Kashani M, Akbari M, Sarvizadeh M. Machine learning for Parkinson's disease: a comprehensive review of datasets, algorithms, and challenges. NPJ Parkinsons Dis 2025;11(1):187.

[21] Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 2021;11(1):3254.

[22] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Adv Neural Inf Process Syst 2014;27.

[23] Touati H, Thabet R, Fontanili F, Cleostrate MH, Pruski M, Cufi MN, Lamine E. AI-driven approach for creating and evaluating a synthetic dataset for medication errors. J Biomed Inform 2025;104889.

[24] Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633 2017.

[25] Vallejo M, Alty J, Smith SL. Deep learning analysis of figure copying tasks for Parkinson's disease detection with GAN-based data augmentation. medRxiv 2025-04, 2025.

[26] Ashrafi N, Schmitt V, Spang RP, Möller S, Voigt-Antons JN. Protect and extend - using GANs for synthetic data generation of time-series medical records. In: 2023 15th Int Conf Qual Multimed Exp (QoMEX), 2023, p. 171-6.

[27] Guo W, Wang J, Wang S. Deep multimodal representation learning: a survey. IEEE Access 2019;7:63373-94.

[28] Teng X, Li X, Wei B. ModFus-PD: synergizing cross-modal attention and contrastive learning for enhanced multimodal diagnosis of Parkinson's disease. Front Comput Neurosci 2025;19:1604399.

[29] Adnan T, Abdelkader A, Liu Z, Hossain E, Park S, Islam MS, Hoque E. A novel fusion architecture for detecting Parkinson's disease using semi-supervised speech embeddings. NPJ Parkinsons Dis 2025;11(1):176.

[30] Lim WS, Fan SP, Chiu SI, Wu MC, Wang PH, Lin KP, et al. Smartphone-derived multidomain features including voice, finger-tapping movement and gait aid early identification of Parkinson's disease. NPJ Parkinsons Dis 2025;11(1):111.

[31] Varghese J, Brenner A, Plagwitz L, van Alen C, Fujarski M, Warnecke T. PADS - Parkinsons Disease Smartwatch dataset. PhysioNet 2024. https://doi.org/10.13026/m0w9-zx22. Version 1.0.0.

[32] Varghese J, Brenner A, Fujarski M, Alen CM, Plagwitz L, Warnecke T. Machine learning in the Parkinson's disease smartwatch (PADS) dataset. NPJ Parkinsons Dis 2024;10(1):9.

[33] Zhao W, Peng X, Wang X, Qi J, Yang Y, Yang P. Multi-scale frequency-aware adversarial network for Parkinson's disease assessment using wearable sensors. In: 2025 IEEE Int Conf Bioinform Biomed (BIBM), 2025, p. 6154-61.

[34] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784 2014.

[35] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 2015.

[36] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Adv Neural Inf Process Syst, 2014, p. 2672-80.

[37] Yang X, Zhang Z, Cui X, Cui R. A time series data augmentation method based on dynamic time warping. In: 2021 Int Conf Comput Commun Artif Intell (CCAI), 2021, p. 116-20.

[38] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.

[39] Breiman L. Random forests. Mach Learn 2001;45(1):5-32.