POLITECNICO DI TORINO

Master's Degree
in Data Science and Engineering

Master's Thesis

Cross-Dataset Classification of Postural Instability in Parkinson's Disease from a Single Wearable Sensor: A Machine Learning Approach

Supervisor
Dr. Luigi Borzì

Candidate
Chiara Maretto

Academic Year 2025/2026

Abstract

Parkinson's Disease (PD) is the most rapidly increasing neurological condition in the world. One of its most debilitating symptoms is postural instability, which causes the loss of balance reflexes and consequently an increased fall risk. Among many scales, it can be clinically evaluated using the MDS-UPDRS (Movement Disorder Society – Unified Parkinson's Disease Rating Scale) Item 3.12 score, which ranges from 0 (stability) to 4 (severe instability).

Many works have explored PD motor symptoms using IMU (Inertial Measurement Unit) sensor data, but postural stability is addressed in far fewer studies compared to gait or tremor. Moreover, existing studies on this topic evaluate their models on a single dataset, due to the high costs of clinical data collection.

This study addresses these gaps following a cross-dataset approach. Three heterogeneous datasets are combined (Kiel, OmniaPark, and WearGait-PD), for a total of 149 subjects performing 2 tasks (stance and walk). For all studies, only data collected from a single IMU on the lower back (accelerometer and gyroscope) were leveraged in this work and MDS-UPDRS Item 3.12 scores were used as labels.

A pre-processing pipeline harmonizes the datasets, followed by a windowing strategy that addresses class imbalance. Latent features are then extracted from windows using an encoder architecture. Specifically, a supervised CNN-GRU (Convolutional Neural Network-Gated Recurrent Unit) classifier and a self-supervised masked autoencoder are trained and compared. Later, latent statistics are derived to obtain a patient-level aggregation and an additional set of handcrafted clinical features is concatenated. This final vector of 46 features is fed into a binary classifier: patients whose MDS-UPDRS Item 3.12 is 0 are labeled as stable, all other values (1,2,3,4) as unstable. Several machine learning models are then compared: Random Forest, Gradient Boosting, Support Vector Machine and Logistic Regression.

Additionally, two domain adaptation strategies, Correlation Alignment (CORAL) and mean-shift Maximum Mean Discrepancy (MMD), are applied to the resulting feature vectors and a dedicated domain classifier is used to quantify domain shift before and after adaptation.

Without domain adaptation, the best models reach a ROC-AUC (Receiver Operating Characteristic Area Under the Curve) of 0.85 for both the supervised classifier and the masked autoencoder encoder. After domain adaptation, performance improves substantially: with CORAL the best configuration reaches an AUC of 0.97. A key finding is that the self-supervised masked autoencoder achieves performance comparable to the supervised encoder, suggesting that unlabeled IMU

data can be effectively leveraged. The domain classifier experiment shows that baseline features are strongly domain-separable (accuracy 0.88–0.97), while both CORAL and MMD effectively reduce this shift. CORAL removes domain information more aggressively than MMD, but both methods improve task performance over the baseline. These results demonstrate that cross-dataset postural instability classification in PD is feasible with a single lumbar sensor, and that feature-level domain adaptation is an effective strategy to improve generalization across heterogeneous clinical datasets.

Contents

List of Tables

List of Figures

Acronyms

AE Autoencoder.

BBS Berg Balance Scale.

BESS Balance Error Scoring System.

CI Confidence Interval.

CNN Convolutional Neural Network.

COP Center of Pressure.

CORAL CORrelation ALignment.

DBS Deep Brain Stimulation.

DL Deep Learning.

FOG Freezing Of Gait.

GBM Gradient Boosting Machine.

GRU Gated Recurrent Unit.

H&Y Hoehn and Yahr.

HAR Human Activity Recognition.

HC Healthy Control.

IMU Inertial Measurement Unit.

LB Lower Back.

LLM Large Language Model.

LOSO Leave-One-Subject-Out.

LR Logistic Regression.

LSTM Long Short Term Memory.

MDS-UPDRS Movement Disorder Society–Unified Parkinson's Disease Rating Scale.

MEMS Microelectromechanical Systems.

ML Machine Learning.

MMD Maximum Mean Discrepancy.

MoCA Montreal Cognitive Assessment.

MSE Mean Square Error.

PCA Principal Component Analysis.

PD Parkinson's Disease.

RBF Radial Basis Function.

ReLU Rectified Linear Unit.

RF Random Forest.

RFE Recursive Feature Elimination.

RMS Root Mean Square.

ROC-AUC Receiver Operating Characteristic Area Under the Curve.

SSL Self-Supervised Learning.

SVM Support Vector Machine.

TUG Timed Up and Go.

UMAP Uniform Manifold Approximation and Projection.

Chapter 1

Introduction

1.1 Parkinson's Disease

Parkinson's Disease (PD) is the most rapidly increasing neurological condition in the world [42]. It was first described in detail in 1817 in James Parkinson's Essay on the Shaking Palsy [41], in which he characterized six cases of patients that had developed a disorder of tremor, slowed movement, and gait disturbance. Many other non-motor symptoms, such as mood disturbance, sleep dysfunction, autonomic dysfunction, cognitive deficits, dementia, and neuropsychiatric symptoms can manifest.

PD is pathologically defined by loss of nigrostriatal dopaminergic innervation, although neurodegeneration also involves cells located in other regions of the neural network. Other processes, such as mitochondrial dysfunction, defective protein clearance mechanisms, and neuroinflammation, have also been linked to PD, but the way these factors interact is still not fully understood. Macroscopically, the brain of a PD patient is characterized by atrophy of the frontal cortex and, in some cases, ventricular dilation. Exact causes remain unknown but both genetic and environmental factors have been demonstrated to play a role. As an example, cigarette smoking and caffeine consumption counterintuitively reduce the risk of developing PD. On the other hand, pesticides, herbicides, and heavy metals exposure has been shown to increase the risk of developing the disease. Moreover, although PD is generally an idiopathic disorder, there is a minority of cases $(10 - 15\%)$ that report a family history [35].

The disease is very common and increasingly spreading, with 0.13 million incident cases worldwide in 2021, which increased by 220.07% compared to 1990 [36]. Today, in Italy, about 250,000 people are affected by PD, of which 10% are under 50 [42].

It is characterized by four main motor symptoms: bradykinesia, rest tremor,

rigidity, and postural instability [35, 55]. Bradykinesia, defined as a generalized slowing of voluntary movement, is the most limiting feature, as it progressively affects the patient's ability to perform everyday tasks such as writing, dressing, and walking [55]. Rest tremor, typically occurring at 4–6 Hz and most visible in the hands, diminishes when the patient intentionally moves the affected limb. Moreover, it is absent in approximately one-third of cases and should not be confused with essential tremor, which behaves in the opposite manner [35]. Rigidity refers to an abnormal increase in muscle tone that resists passive joint movement uniformly in all directions. It produces the typical "cogwheel" phenomenon, a type of stiffness in which a limb reacts with jerks during attempted movement [55]. Postural instability emerges later in the disease course and occurs in 16% of patients [28, 45]. It is caused by a failure of automatic postural reflexes: patients become unable to correct unexpected perturbations to their balance. As the disease progresses, postural instability worsens and often leads to falls [7, 30], which occur in nearly 60% of PD patients [43], and about 75% of the total hospitalizations in PD patients are attributed to falls or fractures [15]. In the advanced stages, gait may also be affected by the disease.

Beyond motor signs, PD involves a variety of non-motor symptoms that affect the large majority of patients and, in many cases, precede the motor ones [55]. This pre-diagnostic window is referred to as the prodromal phase and is characterized by symptoms such as loss of smell, sleep disorders, and constipation. These symptoms have been linked to early neurodegeneration in extranigral structures before dopaminergic cell loss in the substantia nigra becomes clinically evident [35, 55]. Over the course of the disease, patients sometimes also develop orthostatic hypotension, dysphagia, depression, anxiety, cognitive impairment, and hallucinations [35]. These non-motor symptoms contribute substantially to the reduction in quality of life and represent a major driver of hospitalization and care costs [55].

The diagnosis of PD is based on clinical criteria aimed at assessing the presence of bradykinesia and other symptoms such as rigidity and/or tremor. Despite advances in the fields of neuroimaging, genetics, and biomarkers, the accuracy of clinical diagnosis of PD has been suboptimal, especially in the early stages of the disease or in the presence of signs of atypical parkinsonism. The current standard for the assessment of PD is the clinical examination of patients while performing specific tasks, such as walking or sitting, using semi-quantitative rating scales to measure disease progression.

1.2 Clinical Assessment of Postural Stability

The assessment of postural stability in PD is carried out through a set of standardized clinical scales and tests aimed at evaluating the patient's ability to maintain and recover upright posture. Postural instability is a particularly disabling motor symptom, strongly associated with falls and loss of independence, so its quantification is fundamental both to stage disease severity and to monitor therapy [6, 39]. A brief overview of the most commonly used clinical scales is provided below.

MDS-UPDRS Item 3.12. Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [25] is one of the global standards for evaluation and monitoring of PD. The scale consists of 50 total questions divided into four distinct parts, including evaluations from both clinicians and patients. Part III specifically refers to Motor Examination and postural stability is assessed by Item 3.12 through the Pull Test or Retropulsion Test [25]. The examiner stands behind the standing patient, warns them that a pull is about to be applied, and then delivers a quick, forceful backward tug on the shoulders. The examiner observes and scores the corrective postural response on an ordinal scale from 0 to 4 [25]:

- 0 – Normal: no problems; recovers with one or two steps.
- 1 – Slight: 3–5 corrective steps but recovers unaided.
- 2 – Mild: more than 5 steps but recovers unaided.
- 3 – Moderate: stands safely but shows no postural response; would fall if not caught by the examiner.
- 4 – Severe: very unstable; tends to lose balance spontaneously or with only a gentle pull.

Despite its widespread adoption, the Pull Test has several limitations. First, subjects can prepare for the sudden movement, reducing the apparent severity of their deficit. Second, examiners apply the pull force with variable magnitude and duration, making comparisons difficult. Finally, the rating scale collapses several different responses into a single categorical score, so the test can fail to capture early postural instability [32, 49].

Hoehn and Yahr Scale. The Hoehn and Yahr (H&Y) scale [31] is another standard global staging system for PD and is particularly relevant to postural stability because the emergence of impaired postural reflexes defines Stage 3. The scale ranges from Stage 1 (unilateral involvement only) to Stage 5 (wheelchair- or

bed-bound), with the key postural transition at Stage 3: "bilateral disease; mild to moderate disability with impaired postural reflexes; physically independent" [24, 49].

Reaching Stage 3 has a specific relevance: patients at this stage show increasing MDS-UPDRS motor scores, greater cognitive decline, and significantly higher fall risk [24]. For this reason, the H&Y scale is frequently used together with the Pull Test score as a stratification variable in wearable sensor studies of postural instability [6, 49].

The H&Y scale has moderate inter-rater reliability but it has two main limitations for postural assessment. First, it focuses primarily on lower-limb and balance impairments and, second, a single scoring stage covers a wide range of severity [24].

Berg Balance Scale. The Berg Balance Scale (BBS) [5] is a measure of functional balance, divided into 14 items. Each item evaluates a specific functional balance task (e.g., sitting-to-standing, turning 360°, or single-leg stance) using increasingly difficult stances and visual conditions. Items are scored from 0 (unable to perform) to 4 (normal performance), resulting in a total score in the range 0–56.

The BBS has been validated in PD patients, showing good reliability and strong correlations with other balance and disability measures [20]. Scores below 45 are generally associated with an increased risk of falling [20].

Table 1.1 summarizes the three scales described in this section. They address different aspects of postural instability: the Pull Test evaluates postural reflexes, the H&Y scale contextualizes instability within disease progression, and the BBS quantifies functional balance performance. The main common limitations are being ordinal, observer-dependent, and assessed at a single time point in a controlled setting.

Scale	**Domain**	**Range**	**Limitations**
MDS-UPDRS Item 3.12	Postural reflex (Pull Test)	0–4	Subjects can anticipate the pull; variable examiner force; insensitive to early instability [32, 49]
Hoehn & Yahr	Global staging	1–5	Focused on lower-limb impairment; wide severity range within a single stage [24]
Berg Balance Scale	Functional balance	0–56	Scores <45 associated with increased fall risk; assessed in a controlled setting [20]

Table 1.1: Clinical scales for postural stability assessment in PD.

1.3 Technological Approaches to Postural Stability Assessment

A variety of instrumental technologies have been developed to perform the clinical assessment of postural stability.

First, camera-based systems assess posture and movement, capturing either two-dimensional angles from standard video or three-dimensional pose from depth cameras. A photograph or short video clip is used to measure trunk flexion angles with respect to anatomical landmarks [3, 54]. This approach has several advantages: patients are not required to wear any kind of device and it can capture dynamic changes during prolonged or walking tasks [3]. Nonetheless, it requires adequate lighting and zoom conditions and it is computationally expensive to process [3].

Second, optoelectronic systems use multiple calibrated infrared cameras (typically 6 or more) to track the 3D coordinates of passive or active reflective markers placed on the subject's body segments. A biomechanical model then reconstructs full-body kinematics from the recorded marker trajectories [2, 17]. They are considered the gold standard for motion analysis [2, 17], but are very costly, require controlled laboratory settings and are not deployable for home monitoring [2, 23]. Third, force platforms are an established method for assessing postural sway: they measure reaction forces and moments applied by the feet, from which the trajectory of the Center of Pressure (COP) is computed. COP displacement is then used to derive parameters such as total sway path length, mean sway velocity, sway area, and frequency-domain features [6, 23]. Force platforms are widely used in both research and clinical settings and do not require patients to wear any hardware [6, 23], but they are heavy, expensive, and cannot assess postural responses during motion or activities of daily living [23, 47].

Finally, Inertial Measurement Units (IMUs) are wearable sensing devices capable of measuring the motion of a body in three-dimensional space. A complete IMU unit typically integrates three types of sensors: a triaxial accelerometer, a triaxial gyroscope, and, sometimes, a magnetometer. The accelerometer measures linear acceleration in a sensor-fixed frame, capturing both motion dynamics and the gravitational component. The gyroscope measures angular velocity around three orthogonal axes, providing orientation information. The magnetometer measures the local magnetic field vector and is used to correct heading drift in the orientation estimate [23].

Modern IMUs are built on Microelectromechanical Systems (MEMS) technology, which enables the fabrication of microscopic mechanical and electronic components on a single chip [18]. The microscopic scale of MEMS components is used for wearable devices with small form factor, low weight, low power consumption,

and low unit cost. Such characteristics make them perfectly suited for applications in both clinical and everyday monitoring contexts [18, 23].

Sensors can be secured to virtually any body segment using elastic straps or Velcro bands, and the number and location of sensors are determined by the specific clinical question [23, 56]. In PD research, the lower back at the lumbar level is one of the most widely adopted sites for postural and gait assessment, while multi-sensor configurations additionally include the chest, shanks, ankles, and wrists [8, 49, 56].

Moreover, the integration of wireless data transmission into IMU systems enables continuous monitoring of patients in daily living conditions. On the contrary, laboratory measurements taken in a hospital setting may not reflect the patient's typical motor function [18, 47].

Table 1.2 summarizes and compares the principal technologies for postural stability assessment. Compared with camera-based systems, optoelectronic setups, and force plates, IMUs have higher portability, lower cost, and ability to capture postural data across a wide range of tasks and environments. These properties make IMUs particularly well-suited to continuous, long-term monitoring of postural instability in PD, which is the focus of this work.

Feature	**Video / Camera**	**Optoelectronic**	**Force Plate**	**IMU**
Cost	Low–Med	High	Med–High	Low
Portability	Limited	None	None	High
Ambulatory use	No	No	No	Yes
Contact-free	Yes	Yes	Yes	No
3D kinematics	Partial	Full	Indirect	Yes
Dynamic tasks	Yes	Yes	Partial	Yes
Home monitoring	No	No	No	Yes
Setup complexity	Medium	High	Low	Low

Table 1.2: Comparison of principal technologies for postural stability assessment.

1.4 Motivation and Challenges

Currently, the combination of technologies such as wearable devices and artificial intelligence offers several possibilities in medical diagnostics, and in particular for the evaluation of many PD symptoms. The use of these technologies can help overcome the drawbacks of traditional monitoring, such as the lack of objective evaluation.

Machine Learning (ML) and Deep Learning (DL) have been increasingly applied to wearable IMU data for the objective assessment of motor symptoms in

PD. Classical ML approaches, such as Support Vector Machine (SVM) and Random Forest (RF), work on features extracted from inertial signals and have shown promising results in tasks including gait analysis, tremor quantification, and postural stability scoring [44, 47]. More recently, DL architectures such as Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), and their combinations have shown the ability to learn relevant representations directly from raw signals and have achieved competitive performances in several PD motor assessment tasks [33, 47].

Despite this progress, it is still difficult to use these findings in actual patient care. First, even if some datasets for PD motor assessment are available, they are usually limited in size and diversity, because collecting data is expensive and scoring requires expert neurologists [1, 4, 11, 26, 46]. Moreover, strict data privacy regulations limit sharing across institutions [1]. As a consequence, most models are trained and evaluated on a single and often small dataset, while only few leverage more than 100 subjects [47]. Second, a consistent cross-dataset generalization gap has been documented: models that achieve high performance within a single dataset show significantly lower accuracy when tested on data collected with different sensor configurations, body placements, or recording protocols [48]. Third, postural stability has received considerably less attention than other PD motor symptoms such as gait and Freezing Of Gait (FOG): a recent survey found that only a few studies have specifically targeted the pull test using IMU data, compared to the larger number of works on gait and tremor analysis [47, 49]. This work addresses these challenges by developing a cross-dataset approach for postural instability assessment.

Chapter 2

Related Work

2.1 Wearable Sensor Data for Motor Symptoms Assessment

Wearable IMUs have become one of the most widely used tools for the objective assessment of motor symptoms in PD, since they are suitable for both clinical and home settings, where continuous monitoring of patients during daily activities is increasingly needed [18, 53]. These sensors can provide a detailed description of body movement, which can be directly related to the motor impairments of PD [23]. The sensor placement depends on the motor symptom of interest. For gait analysis, sensors are typically attached to the feet, shanks, or lower back. For postural assessment, the lower back at the lumbar level is the most widely used location, as it lies close to the body's center of mass and captures the overall trunk movement [23]. For tremor and bradykinesia, wrist-mounted sensors are more common. Multi-sensor configurations cover several body segments and can therefore provide additional information on motor impairment, but complexity increases and patient compliance reduces [53]. One important limitation of clinical data collected with IMUs is the variability introduced by differences in sensor type, sampling frequency, and recording protocol across studies and institutions. These differences make it difficult to combine data from multiple sources or to apply a model trained on one dataset directly to another [48]. This is a fundamental challenge that motivates the use of domain adaptation methods, discussed later in this chapter.

2.2 Feature Extraction from Inertial Signals

The majority of IMU-based studies on PD in literature relies on handcrafted features extracted from raw signals before training a classifier. These features

can be divided into three main categories: time domain, frequency domain, and amplitude-power domain. Time domain features are computed directly from the signal samples. Common examples include the Root Mean Square (RMS) of the acceleration, signal range, peak and trough values, standard deviation, skewness, kurtosis, and entropy-based measures [23, 49]. Frequency domain features describe the distribution of signal power across frequency bands. For postural assessment, slow sway frequencies between 0.1 and 0.5 Hz are particularly relevant [23]. Features such as peak power frequency, centroidal frequency, and the ratio of power in specific bands to total power have been shown to discriminate between different levels of postural impairment [49]. As an example, Smith et al. [49] computed a total of 60 features from accelerometer signals recorded during the pull test, finding that frequency-based features achieved the highest classification performance, with two features (chest-to-lumbar maximum acceleration ratio and 8–12 Hz to total acceleration power ratio) reaching a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.95.

Wearable sensors typically produce a significant number of potential features, not all of which are useful for a given clinical task. This makes feature selection an important preprocessing step. Common approaches include filter methods based on statistical tests, wrapper methods such as sequential forward selection, and dimensionality reduction with Principal Component Analysis (PCA) [50]. Sotirakis et al. [50] showed that a random forest regressor, combined with PCA, could track the longitudinal progression of PD motor symptoms from walking and standing tasks, using features extracted from six IMUs.

Despite the success of handcrafted features, their definition and extraction is time-consuming and requires domain expertise. As an alternative, DL approaches have been investigated to learn relevant representations directly from the raw signal.

2.3 Machine Learning and Deep Learning for Motor Symptoms

ML and DL methods have been applied to IMU data for a wide range of PD motor assessment tasks, including gait analysis, tremor quantification, FOG detection, and bradykinesia scoring. Sigcha et al. [47] reviewed and analysed 69 studies that combined wearable sensors and DL for PD. Inertial sensors represented the most widely adopted modality and CNNs were employed in more than half of the studies. More than 50% of the works focused on gait impairments and tremor, while postural instability, bradykinesia, and rigidity received considerably less attention. This distribution reflects the relative availability of labeled data and the difficulty of instrumenting certain clinical tests. Classical ML approaches, such as SVMs and

RFs, are still widely used when working with handcrafted features, since they are simple to train, interpretable, and effective when the number of samples is limited [47, 50]. However, they require manual feature design and may not capture the full complexity of the inertial signal. On the other hand, DL models learn hierarchical representations directly from the raw data. CNNs are particularly effective at extracting local temporal patterns from time series, while recurrent architectures such as Long Short Term Memory (LSTM) and GRU are leveraged to capture longer temporal dependencies. Hybrid architectures that combine convolutional and recurrent layers have shown strong performance on balance and activity recognition tasks. Kim et al. [33] proposed a model with two 1D-CNN heads and one GRU head for automatic scoring of the Berg Balance Scale from IMU signals. They achieved 98.4% accuracy across 14 balance tasks. The main finding was that mixing convolutional and recurrent blocks resulted in better performance than either architecture used alone. This suggests that both local feature extraction and long temporal dependencies contribute to classification performance. The review by Sigcha et al. [47] also highlighted a recurrent problem: nearly all studies train and evaluate their models on a single dataset, and when cross-dataset evaluation is performed, performance drops significantly.

2.4 Postural Instability: An Underexplored Motor Symptom

Despite its clinical importance, postural instability is one of the least studied motor symptoms of PD in the wearable sensor literature. A search of the PubMed database conducted by Smith et al. [49] found more than 75 studies on gait analysis, more than 50 on sway or tremor, but only 5 studies that directly used IMU data to characterise PD using the pull test. This imbalance is likely due to the nature of the test: gait and tremor are assessed using continuous signals during everyday activities, while the pull test is very quick, depends on the examiner and needs to be performed in a clinical setting.

Early work showed the potential of wearable sensors for postural assessment, without focusing on the pull test. Mancini et al. [37] demonstrated that a single trunk sensor could detect signs of postural instability in PD even before they were clinically apparent. Palmerini et al. [40] showed that a small set of accelerometer features during quiet standing was sufficient to stratify patients by level of postural impairment. A systematic review covering 47 studies on wearable balance assessment [23] confirmed that the lower back is the preferred sensor location, and that sway area, jerk index, and spectral power in slow frequencies are among the most discriminative features for PD populations. The review also noted that virtually no study had validated its model on an external dataset.

More recent work has focused on predicting the MDS-UPDRS postural stability score directly. Borzì et al. [9] used a waist-mounted smartphone during 30 seconds of quiet standing and trained an SVM on over 400 extracted features. The model achieved excellent classification accuracy on the available dataset of 42 patients, but no external validation was performed. Smith et al. [49] focused on the pull test, leveraging data from 6 sensors and extracting 60 features to separate Hoehn and Yahr Stage II from Stage III in 30 patients. Two features (chest-to-lumbar maximum acceleration ratio and 8–12 Hz to total acceleration power ratio) achieved a ROC-AUC of 0.95. The study also showed that classification performance did not worsen when the model was applied to scores from two different examiners, addressing one source of variability. However, the authors acknowledged that multi-site validation on larger cohorts is necessary.

2.5 Self-Supervised Learning for Sensor Signals

One of the main difficulties in applying DL to clinical IMU data is the lack of labeled examples. Collecting labeled data in PD research requires skilled neurologists to perform and score clinical tests, which is expensive and time-consuming. Self-Supervised Learning (SSL) addresses this by learning useful representations from large amounts of unlabeled data before fine-tuning on a small labeled set. A widely used family of SSL methods for time series is based on masked reconstruction. The model is trained to reconstruct portions of the signal that have been randomly hidden. This forces the encoder to learn a compact representation of the signal structure that can then be transferred to the target classification task. Cheng et al. [14] proposed MaskCAE, which applies sparse convolution to process only the visible parts of the signal during pre-training and then it reconstructs the masked segments using a decoder. MaskCAE outperformed contrastive learning baselines on three public Human Activity Recognition (HAR) datasets and was shown to be highly computationally efficient. Miao et al. [38] extended this idea to multi-sensor settings with the STMAE model. It applies two stages of masking: one at the sensor channel level to simulate missing devices, and one at the time step level. The model learns to recover both the spatial structure across sensors and the temporal dynamics within each channel. On four public datasets, STMAE outperformed other SSL approaches, particularly in settings where some sensors were absent at test time. This robustness to missing channels is especially relevant for clinical applications, where sensor failures or changes in recording configuration are common.

2.6 Domain Adaptation for Cross-Dataset Generalisation

When a model is applied to data from a different source than the one used for training, performance typically decreases. This happens because of differences in sensor hardware, sampling rates, sensor placement, recording protocols, and patient populations, which are reflected in the statistical distributions of training and test features, a problem known as domain shift [48]. Domain adaptation methods aim to reduce this gap without requiring labeled data from the target domain. Two well-established approaches are CORrelation ALignment (CORAL) and Maximum Mean Discrepancy (MMD). CORAL [51] aligns the second-order statistics of the source and target distributions by applying a linear transformation to the source features so that their covariance matrix matches the target one. Sun and Saenko [52] extended this approach to deep networks. They added a differentiable CORAL loss that can be minimized along with the classification loss during training to align features. MMD [27] measures the difference between two distributions by comparing their mean embeddings in a reproducing kernel Hilbert space. In its simplest linear form, it corresponds to a mean-shift that brings the source feature distribution closer to the target. MMD is computationally simpler and generally more stable when the number of samples per domain is small. Sigcha et al. [48] specifically addressed cross-dataset generalization in FOG detection. Despite standardizing the signal scale and sampling rate across three datasets, a model trained on one source generalized poorly to the others. Their analysis showed that differences in sensor type, body placement, and recording conditions all contributed to the gap, and that simply training on more data was not sufficient to close it. This underlines the need for explicit alignment methods. While CORAL and MMD have been extensively studied in computer vision and natural language processing, their application to cross-dataset IMU-based postural stability classification in PD still needs to be explored.

Table 2.1 summaries the studies related to this work, covering IMU-based approaches for the objective assessment of postural stability in PD. For each study, the table reports the dataset used (number of subjects and amount of data), the sensor configuration (number, type and body placement), the experimental protocol (tasks and conditions), the clinical target scale, the pre-processing and modeling pipeline, the validation strategy, and the main performance metrics.

Reference	Dataset	Sensors	Protocol	Target	Pre-processing & Method	Validation	Results
Borzì et al. 2020 [9]	42 PD + 7 HC; single site (Turin); 30 s per subject	1 smartphone (triaxial acc + gyr); LB at L3–L5	Quiet stance, 30 s; daily ON state; hospital outpatient visit	MDS-UPDRS Item 3.12 (score 0–3)	414 time/frequency features; Pearson selection → 8 features; Linear SVM, two-layer cascade	LOSO	HC vs. PS: Acc = 100%; multi-class SVM: Acc = 70.1%, avg F1 = 67.3%
Smith et al. 2025 [49]	30 PD (H&Y II vs. III); single site (Madrid); 3 pull tests per subject at 50 Hz	6 IMUs (acc); chest, lumbar, shanks, feet; anteroposterior axis only	Pull test (3 reps); OFF medication; 2 examiners	H&Y stage II vs. III	Bandpass 0.55–20 Hz; 60 features (time, frequency, power); age correction; Mann–Whitney U + ROC-AUC	Statistical feature ranking	Best AUC = 0.95 (chest-to-lumbar max acc ratio; power ratio 8–12 Hz vs. total)
Kim et al. 2021 [33]	53 subjects (stroke rehab + healthy); single site (Inha Univ. Hospital, Korea)	1 IMU (triaxial acc + gyr); sternum/trunk	14 BBS tasks; supervised by rehab therapist	BBS item score (0–4) per task	Down-sampling to 20 Hz; oversampling augmentation; stacking ensemble: 2×1D-CNN + 1×GRU with dense meta-learner	10-fold cross-validation	Avg acc = 98.4% across 14 BBS tasks
Ghislieri et al. 2019 [23]	Systematic review: 47 studies; multiple populations incl. PD	IMU (acc, gyr, mag); most common: 1 sensor at L5 (42.6% of studies)	Quiet stance: eyes open/closed; foam/firm surface; tandem/single-leg	Clinical scores (BBS, BESS) or force plate COP	Lowpass 0.5–10 Hz; RMS, jerk index, range, centroidal frequency, frequency dispersion	vs. force plate (10 studies) or clinical scores (7 studies)	Best PD discriminators: jerk index, sway amplitude (time domain); frequency dispersion, centroidal frequency (freq. domain)

Table 2.1: Summary of related work on IMU-based postural stability assessment in PD.
Abbreviations: PD = Parkinson's Disease; HC = Healthy Controls; PS = Postural Stability; LB = Lower Back; acc = accelerometer; gyr = gyroscope; mag = magnetometer; MDS-UPDRS = Movement Disorder Society Unified Parkinson's Disease Rating Scale; H&Y = Hoehn & Yahr; BBS = Berg Balance Scale; BESS = Balance Error Scoring System; COP = Centre of Pressure; RMS = Root Mean Square; SVM = Support Vector Machine; CNN = Convolutional Neural Network; GRU = Gated Recurrent Unit; ROC-AUC = Receiver Operating Characteristic Area Under the Curve; LOSO = Leave-One-Subject-Out; Acc = Accuracy; F1 = F1 score.

2.7 Contributions and Thesis Organization

The literature reviewed above shows three main gaps. First, existing studies on IMU-based postural stability in PD use a single dataset and their models are not tested on data from a different clinical site [9, 23, 49]. Second, handcrafted signal features and deep learned representations have been used separately, but no study has combined them at the patient level for this specific task. Third, domain adaptation methods have not been applied to cross-dataset postural stability classification in PD. This work addresses these gaps, and its main contributions are summarised in the following:

- A cross-dataset approach that combines three datasets. First, a pre-processing pipeline is used to create a uniform format. It includes signal filtering, axes alignment, resampling to a common frequency, and an adaptive windowing strategy to overcome class imbalance.
- A systematic comparison of two encoder architectures for learning patient-level representations: a supervised CNN-GRU classifier and a self-supervised masked autoencoder. Both produce an 8-dimensional latent vector per time window, which is then aggregated across all windows of a patient.
- A patient-level feature representation that combines the 32-dimensional latent block with 14 handcrafted clinical features (8 from quiet stance, 6 from walking), for a total of 46 dimensions per patient.
- An ablation study that tests all combinations of four classifiers (RF, Gradient Boosting Machine (GBM), SVM, Logistic Regression (LR)), four feature subsets (handcrafted only, latent only, combined, combined with feature selection), and three variants of domain adaptation (no adaptation, CORAL, mean-shift MMD), across two encoder architectures.
- A domain classifier that predicts the source dataset given the feature vector of a patient. This gives a direct quantitative measure of domain shift.
- An ordinal analysis of probabilities from binary classification, to understand whether they convey ordinal information about the MDS-UPDRS scale.

The rest of this work is structured as follows. Chapter 3 describes the datasets, pre-processing pipeline, and modeling choices. Section 3.1 introduces the three clinical datasets (WearGait-PD, Kiel, and OmniaPark) and details their common characteristics that make a cross-dataset approach feasible. Section 3.2 covers the harmonisation steps, including axes alignment, filtering, resampling, and the adaptive windowing strategy used to address class imbalance. Section 3.3 describes the

two encoder architectures, the patient-level aggregation of latent statistics and the handcrafted features. Section 3.4 presents the four classical ML classifiers compared in the ablation study and the feature selection procedure. Section 3.5 discusses domain shift and the adaptation strategies implemented to address it (CORAL and MMD). Section 3.6 defines the training procedure the evaluation metrics employed in the work, and Section 3.7 describes the ordinal severity estimation.

Chapter 4 reports the experimental results. Section 4.1 analyses domain shift before and after adaptation using the domain classifier. Section 4.2 presents the full ablation study across encoder architectures, feature subsets, domain adaptation variants, and classifiers. Section 4.3 reports the results of the ordinal estimation analysis and, finally, Section 4.4 compares the findings of this work with previous related studies.

Chapter 5 interprets the results and situates them within the existing literature. It discusses the cross-dataset generalization of the approach, the motivation for the hybrid architecture, the comparison between the supervised and self-supervised encoders, and the contribution of each feature subset. It then analyses the distinct effects of CORAL and MMD on domain shift and classification performance, provides a clinical interpretation of the model predictions, and concludes by outlining the main limitations of the study.

Chapter 6 summarizes the main conclusions and outlines directions for future research.

Chapter 3

Materials and Methods

This chapter describes the complete methodology, whose overall structure is summarized in Figure 3.1. Three heterogeneous clinical datasets (WearGait-PD, Kiel, and OmniaPark) are introduced in Section 3.1, retaining only PD patients and the accelerometer and gyroscope signals of a single lower-back IMU. Because the datasets differ in hardware, sampling frequency, and axis conventions, Section 3.2 details the pre-processing pipeline that harmonizes the data, including unit and axes standardisation, outlier removal, filtering, resampling, and an adaptive windowing strategy designed to handle class imbalance. Each window is then converted into a patient-level representation in Section 3.3. Latent features are extracted by two alternative encoders, a supervised CNN-GRU classifier and a self-supervised masked autoencoder, aggregated across windows, and concatenated with a set of handcrafted clinical descriptors. Section 3.4 presents the binary classification of postural instability, comparing four classical ML models together with the feature selection procedure. Section 3.5 describes the domain classifier used to quantify domain shift and explains the adaptation methods employed to counter it. Section 3.6 defines the training procedure and the evaluation metrics. Finally, Section 3.7 presents the exploratory ordinal estimation of severity derived from the binary probabilities.

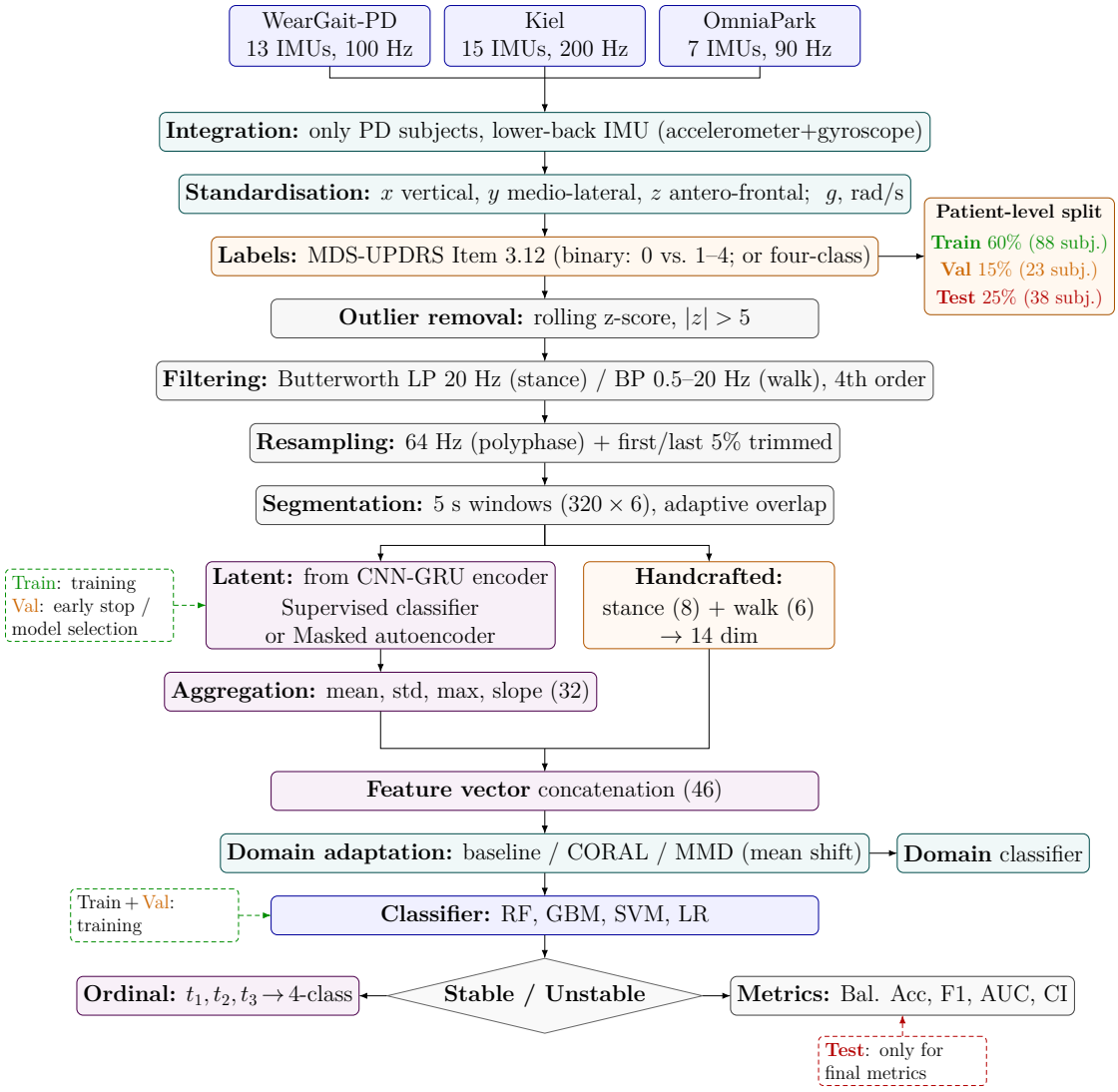


Figure 3.1: Overview of the processing pipeline. Abbreviations: IMU = Inertial Measurement Unit; PD = Parkinson's Disease; MDS-UPDRS = Movement Disorder Society Unified Parkinson's Disease Rating Scale; LP = low-pass; BP = band-pass; CNN-GRU = Convolutional Neural Network–Gated Recurrent Unit; std = standard deviation; CORAL = Correlation Alignment; MMD = Maximum Mean Discrepancy; RF = Random Forest; GBM = Gradient Boosting Machine; SVM = Support Vector Machine; LR = Logistic Regression; ML = Machine Learning; Bal. Acc = Balanced Accuracy; F1 = F1 score; AUC = Area Under the Curve; CI = Confidence Interval; Train/Val/Test = training/validation/test split.

3.1 Data

This section details the three datasets used in this work, describing sensors, subjects and acquisition procedures.

WearGait-PD. WearGait-PD is a large and open-access dataset designed for the development of algorithms for gait and balance assessment in PD [1]. It was collected at three clinical sites in the United States and includes data from 100 people with PD (35 female, 65 male; mean age 67 ± 8 years) and 85 Healthy Controls (HCs). Participants performed several gait and balance tasks, including self-paced walking, hurried-pace walking, walking through a mock doorway, and a free-walk task. Data were collected at 100 Hz from 13 IMUs (triaxial accelerometer, gyroscope, magnetometer, and orientation) placed on the head, sternum, wrists, lower back, thighs, shanks, ankles, and dorsal feet. A pressure-sensitive walkway was used to record foot contact timing and pressure during walking. Two expert reviewers annotated video frames, indicating general activities, FOG episodes, and clinically relevant events such as start hesitation. Clinical data, such as MDS-UPDRS scores, modified H&Y staging, medication state, and deep brain stimulation status are provided for each participant. The dataset is publicly accessible via the Synapse platform under a Creative Commons CC BY 4.0 license [34].

Kiel. The Kiel dataset is a mobility dataset designed to support the development and validation of IMU-based algorithms [56]. A total of 167 subjects were enrolled at the University Hospital Schleswig-Holstein (Campus Kiel, Germany): 43 healthy younger adults (18–60 years), 24 healthy older adults (>60 years), 34 patients with PD (mean age 65 ± 11 years), 23 who had a stroke, 21 with multiple sclerosis, 10 with chronic low back pain, and 12 with other neurological movement disorders. Each participant was equipped with at least 15 IMUs (triaxial accelerometer, gyroscope and magnetometer) placed on the head, sternum, upper and forearms, pelvis, thighs, shanks, ankles, and feet, recording at 200 Hz (100 Hz for a subset of 12 subjects). Simultaneously, an optical motion capture system provided kinematic reference measurements. Participants performed a sequence of standardized clinical assessments, such as the Timed Up and Go test, 10-meter walk, and treadmill walking at preferred speed, as well as daily activities. Clinical data include MDS-UPDRS Item 3 scores, H&Y staging, and the Montreal Cognitive Assessment. 21 HC are openly available, while accessing the other groups requires a data sharing agreement with the authors [29].

OmniaPark. The OmniaPark dataset was collected as part of the Objective Monitoring of Axial symptoms in Parkinson's Disease (OMNIA-PARK) study [22],

a multicenter, observational protocol aimed at the development and validation of a technological setup for the objective assessment of axial symptoms in PD [22]. Data were collected at two Italian Movement Disorders Centers (Turin and Rome) from 40 PD patients. In the first phase, participants performed gait and balance tasks in a supervised laboratory setting. Several clinical parameters are included, such as MDS-UPDRS Item 3 and BBS, along with standard demographic information.

Kinematic data were collected using the Movit system (Captiks Srl) [13], a wireless IMU platform providing triaxial accelerometer and gyroscope measurements. Seven IMUs were placed on the ankles, lower back, wrists, spine, and head. Raw signals were resampled to 90 Hz.

Table 3.1 summarizes and compares the main characteristics of the three datasets.

Feature	WearGait-PD [1]	Kiel [56]	OmniaPark [22]
Participants	100 PD, 85 HC	67 HC, 34 PD, 23 stroke, 21 MS, 10 CLBP, 12 other	40 PD
Setting	Three clinical sites (USA)	University hospital (Kiel, Germany)	Two movement disorders centers (Turin & Rome, Italy)
IMU system	Movella/Xsens MTw Awinda	Noraxon myoMOTION	Movit (Captiks Srl)
No. of IMUs	13	$\geq$15 (up to 16)	7
IMU placement	Head, sternum, lower back, wrists, thighs, shanks, ankles, dorsal feet	Head, sternum, upper/forearms, pelvis, thighs, shanks, ankles, feet	Ankles, lower back, wrists, spine, head
Signals	Accel. (g), gyro. (rad/s), mag., orientation	Accel. (g), gyro. (deg/s), mag.	Accel. (g), gyro. (deg/s)
Sampling freq.	100 Hz	200 Hz (100 Hz for 12 subjects)	90 Hz
Main tasks	Self-paced & hurried walking, tandem gait, TUG, balance, mock doorway, free walk	Overground walking (slow/preferred/fast), backwards/sideways walking, TUG, treadmill, sit-to-stand, slalom, activities of daily living	Gait tasks, balance tasks (stand open/closed eyes)
Clinical data	MDS-UPDRS (all parts), modified H&Y, medication state & timing, DBS status, demographics	MDS-UPDRS III, H&Y, MoCA, SARC-F, disease duration, medication, demographics	MDS-UPDRS III (postural stability, posture, arising from chair), H&Y, BBS, disease duration, demographics

Table 3.1: Summary of the three datasets used in this work.
Abbreviations: PD = Parkinson's Disease; HC = Healthy Controls; MS = Multiple Sclerosis; CLBP = Chronic Low Back Pain; IMU = Inertial Measurement Unit; Accel. = accelerometer; gyro. = gyroscope; mag. = magnetometer; TUG = Timed Up and Go; DBS = Deep Brain Stimulation; MDS-UPDRS = Movement Disorder Society Unified Parkinson's Disease Rating Scale; H&Y = Hoehn & Yahr; MoCA = Montreal Cognitive Assessment; SARC-F = Strength, Assistance with walking, Rising from a chair, Climbing stairs, and Falls (questionnaire); BBS = Berg Balance Scale.

3.2 Pre-processing

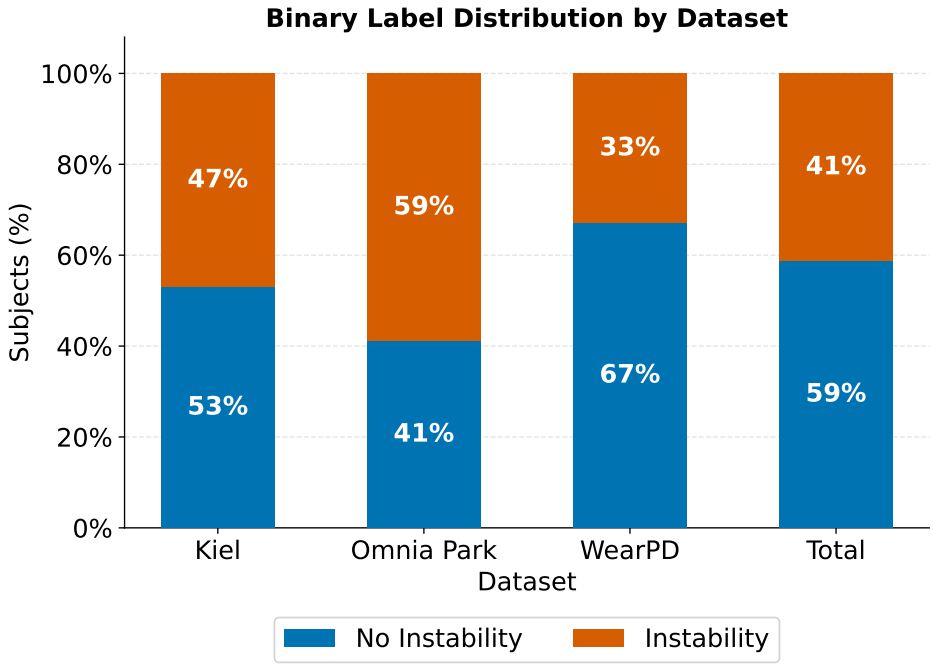
The three datasets used in this work differ significantly in sampling frequency, axes orientation, and measurement units. The pre-processing pipeline was designed to produce uniform characteristics to train a single cross-dataset model.

Dataset Integration. The three datasets described above (WearGait-PD, Kiel, and OmniaPark) share a set of common properties that make a cross-dataset approach feasible. First, only PD patients with both stance and walking tasks available were included in the analysis, for a total of 149 subjects (17 from Kiel, 38 from OmniaPark, and 94 from WearGait-PD). Notably, walking recordings include both gait and turning samples, which are treated equivalently in the following steps. Second, different datasets employed a variable number of IMUs on various body parts. In this work, only the sensor located on the lower back was used, as it represents the most common and clinically motivated placement for balance and gait analysis in PD [23, 37]. Specifically, only the triaxial accelerometer and triaxial gyroscope signals from this sensor were employed, for a total of 6 input channels per time step.

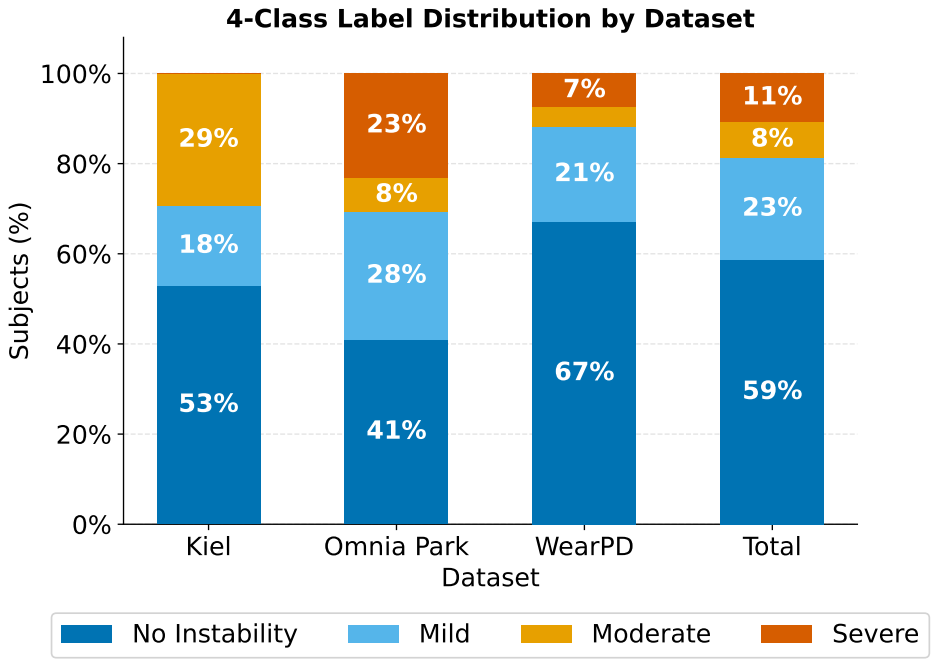
Labels Extraction. The ground-truth label for postural instability was derived from the MDS-UPDRS Item 3.12 score, which represents the clinical outcome of the pull test from 0 to 4 [25]. Two different labeling schemes were adopted. In the binary case, score 0 was mapped to class 0 (no instability), while scores 1, 2, 3, and 4 were mapped to class 1 (instability present). In the four-class scheme, scores 0, 1, and 2 were kept as separate classes (No Instability, Mild, and Moderate), while scores 3 and 4 were merged into a single Severe class since fewer examples were present. Figure 3.2 shows the label distribution across the three datasets and in the merged one, comparing the two schemes. In the binary case, the final distribution is moderately unbalanced, with 59% of subjects belonging to the class 0 and 41% to class 1. The imbalance varies between datasets: WearGait-PD is the most skewed (67% vs. 33%), while Kiel and OmniaPark are more balanced. On the other hand, the four-class distribution is highly unbalanced, with 23% of subjects belonging to the Mild class, and only 8% to Moderate and 11% to Severe.

The demographic composition of the merged dataset is summarized in Figure 3.3. Males are more present in the sample (66%), accordingly to the known higher incidence of PD in men [36]. The age distribution is centered around 63–65 years, with most subjects falling between 60 and 75 years of age.

Axes and Measurement Units Standardization. Data were collected with heterogeneous devices and different axis orientations. First, to standardize the datasets, all signals were converted to common measurement units: g for the

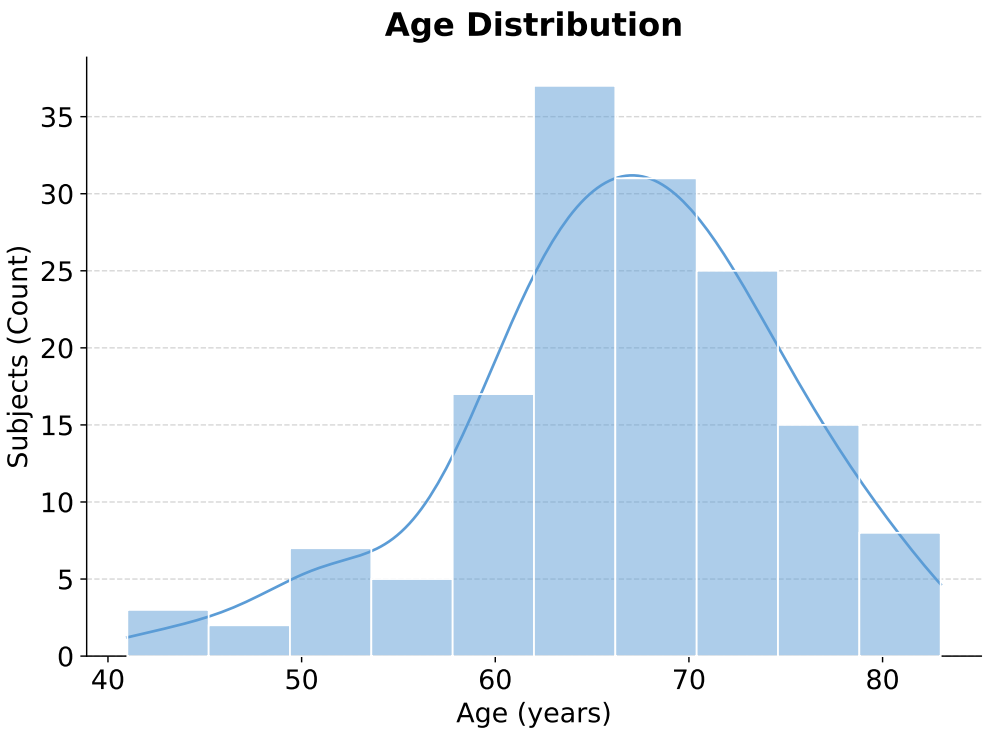


(a) Binary label distribution (score 0 vs. scores 1–4).

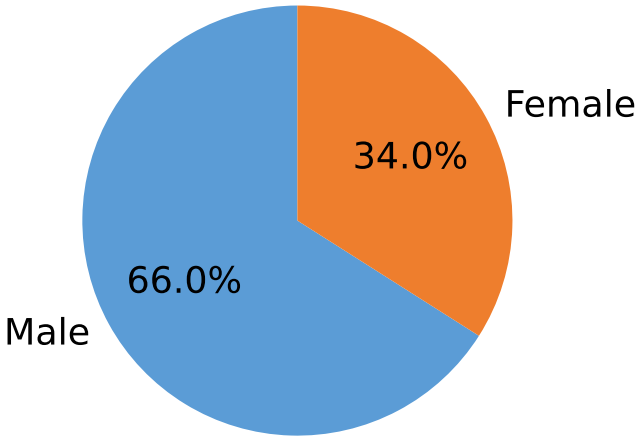


(b) Four-class label distribution (scores 0, 1, 2, and 3–4 merged).

Figure 3.2: Distribution of MDS-UPDRS Item 3.12 labels across the three datasets under the binary (a) and four-class (b) labeling schemes.



(a) Age distribution.



(b) Gender distribution.

Figure 3.3: Age (a) and gender (b) distributions of the merged dataset of 149 participants across the three datasets.

accelerometer and $\frac{rad}{s}$ for the gyroscope. Axes orientations were aligned so that the x-axis corresponds to the vertical direction (positive upward), the y-axis to the medio-lateral direction (positive right) and the z-axis to the antero-posterior direction (positive backward).

Outliers Removal. Outliers removal is performed via a rolling z-score approach. For each inertial signal channel, a local mean μ and standard deviation σ are computed over a centered sliding window of 2 seconds and then the z-score is computed as: $z = \frac{x-\mu}{\sigma}$. Samples whose z-score exceeds the threshold $|z| > 5$ are considered outliers. This strategy eliminates artifacts caused by sensor disconnections or sudden movements while preserving signal shape.

Filtering. Different filtering strategies are applied based on the task. For stance windows, each channel is filtered with a fourth-order, zero-lag Butterworth low-pass filter with cut-off frequency at 20 Hz to eliminate high-frequency noise while retaining low frequency components that may contain useful stability information. Differently, a fourth-order Butterworth band-pass filter between 0.5 Hz and 20 Hz is used for walking windows to remove also low frequency drift and the effect of gravity.

Resampling and Trimming. Polyphase filtering is used to resample all recordings to the target frequency of $f_t = 64$ Hz. This value was chosen as a trade-off between the temporal resolution and computational complexity. Additionally, before segmentation, the first and final 5% of each recording are trimmed to exclude non-representative data and ensure that only relevant time intervals are analyzed.

Adaptive Segmentation. The resampled signals are divided into fixed-length temporal windows of $T_w = 5$s, using an adaptive segmentation strategy [10] designed to address class imbalance across MDS-UPDRS Item 3.12 values. It is worth noticing that the scale ranges from 0 to 4, but classes 3 and 4 were aggregated due to fewer samples, as explained above.

The procedure consists of three phases. First, the baseline number of valid windows producible from each class is computed using a minimum overlap of $\delta_{\min} = 0.5$. The majority class is identified and a target window count is set as $N^* = \lfloor N_{\max} \cdot r \rfloor$. The target ratio is set to $r = 0.85$, meaning that for each class, the goal is to reach a number of samples corresponding to 85% of the total count of the majority class. In the second phase, for each class c the overlap required to reach N^* is derived by inverting the relationship:

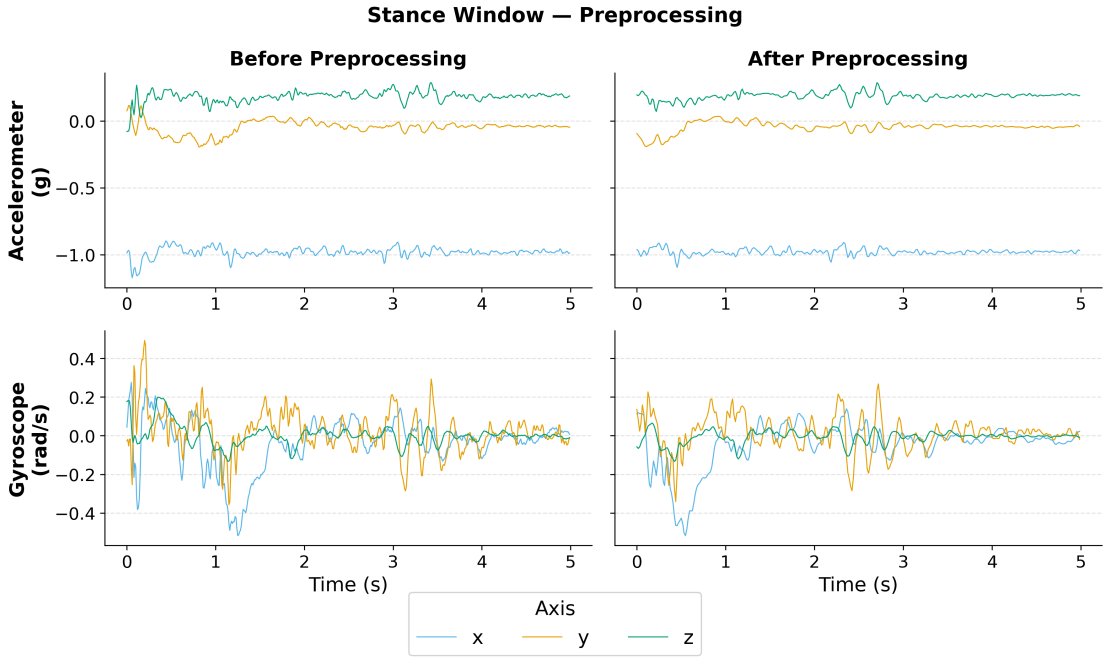
$$\delta_c = 1 - \frac{1 - \delta_{\min}}{N^*/N_c}, \tag{3.1}$$

where N_c is the baseline window count of class c. The resulting value is clipped to the interval $[\delta_{\min}, \delta_{\max}] = [0.5, 0.85]$. In this way, minority classes are assigned a higher overlap, increasing the number of extracted windows. Finally, windows in which fewer than 5% of samples are missing are recovered via cubic polynomial interpolation fitted independently on each channel, otherwise they are discarded. In the end, only subjects who have at least one valid window for both task types (walking and stance) are kept. Each window is represented as a tensor of shape (320,6), corresponding to $T_w \cdot f_t = 5 \cdot 64$ time steps and 6 inertial channels.

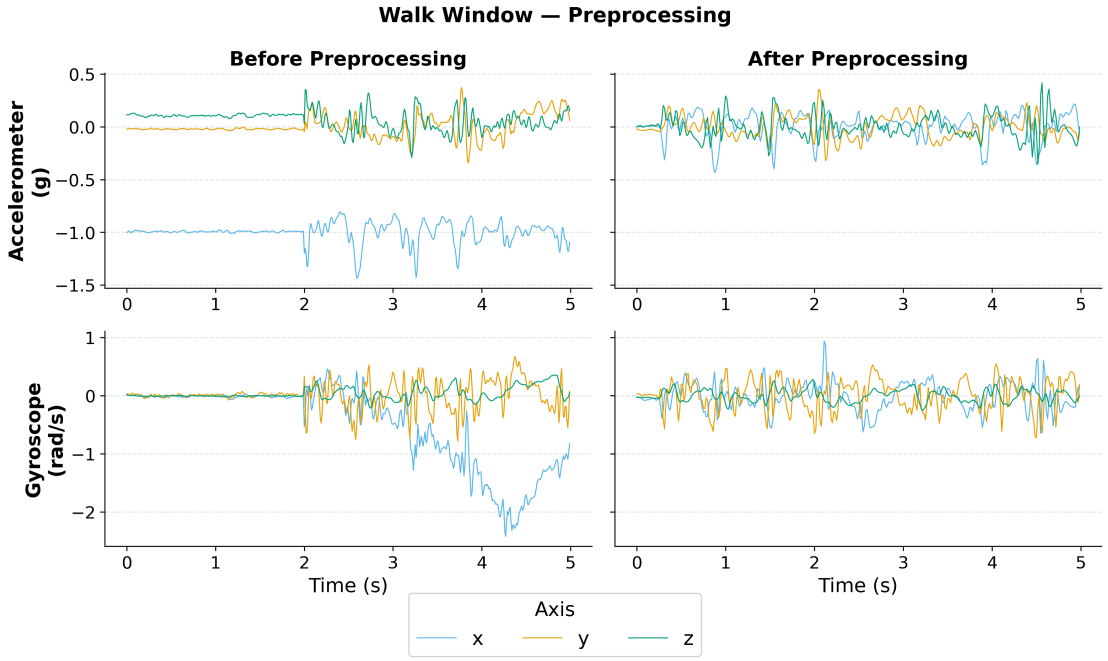
Table 3.2 summarizes the number of windows extracted after the full pre-processing pipeline. Before applying adaptive segmentation, the class distribution was highly skewed: 2109 windows for class 0 but only 373 for class 2. After applying the adaptive overlap strategy, the window counts are substantially more balanced. The final dataset contains 6998 windows across the four classes. Examples of stance and walk windows before and after pre-processing are provided in Figure 3.4. These plots highlight the distinct signal patterns associated with each activity, as well as the efficacy of the pre-processing steps in removing outliers and trimming non-representative initial and final sample.

Class	**Description**	$\boldsymbol{\delta_c}$	**Baseline**	**After Adaptation**
0	No instability	0.50	2109	2165
1	Mild	0.66	1230	1799
2	Moderate	0.85	373	1241
3–4	Severe	0.74	916	1793
Total			**4628**	**6998**

Table 3.2: Effect of adaptive segmentation on window counts per class.



(a) Example of a 5s stance window before and after the complete pre-processing pipeline.



(b) Example of a 5s walk window before and after the complete pre-processing pipeline.

Figure 3.4: Example of a 5s window during stance (a) and walk (b) before and after the complete pre-processing pipeline.

3.3 Feature Extraction

Subsequently, each pre-processed window is fed into an encoder architecture to extract an 8-dimensional latent representation. All windows of a patient are then aggregated using 4 statistics (mean, standard deviation, maximum and slope), resulting in 32 features. Additionally, 14 handcrafted descriptors are extracted: 8 from stance and 6 from walking windows. Finally, all features are concatenated in a 46-dimensional patient vector, which is then fed to a classifier for postural stability assessment.

The dataset is split at the patient level: 60% training (88 subjects), 15% validation (23 subjects), and 25% test (38 subjects) sets, stratified by binary label. The split is performed before any encoder training to prevent information leakage. All windows belonging to the same patient are assigned exclusively to one partition. Prior to feature extraction, each patient's windows are independently z-score normalized across channels to mitigate inter-subject amplitude differences.

3.3.1 Latent Features

Two encoder architectures are evaluated as alternatives for latent features extraction: a supervised CNN-GRU classifier and a self-supervised masked autoencoder. Both share the same backbone structure and produce an 8-dimensional latent vector per window. Both architectures leverage the Adam optimizer with learning rate 10^{-4}, gradient norm clipping at 1.0, and a schedule that halves the learning rate after 10 epochs of patience. Training proceeds for up to 100 epochs with an early stop strategy based on validation accuracy (patience of 20 epochs), model check-pointing on the best validation accuracy, and batch size of 64.

Supervised Classifier

The model takes an input tensor of shape $(T \times C) = (320 \times 6)$, which represents a 5 seconds window, sampled at 64 Hz across 6 channels. The goal is to correctly classify each window according to the MDS-UPDRS Item 3.12 score of the corresponding patient. As explained before, the scale ranges from 0 to 4, but classes 3 and 4 were aggregated.

The complete architecture is shown in Figure 3.5a. Specifically, the first convolutional layer applies 16 filters of kernel size 5, while the second applies 32 filters of kernel size 3, both followed by batch normalization. The GRU layer has 32 hidden units and returns a single vector summarizing the full temporal sequence. A single dropout layer (rate $p = 0.4$) is applied to the GRU output, before the latent dense layer, and only during training, to improve generalisation. The bottleneck dense layer projects to $d = 8$ dimensions, producing the latent vector. All layers use

Rectified Linear Unit (ReLU) activation functions. A final softmax dense layer outputs the final class probabilities for the training task. The model is trained with sparse categorical cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{4}\sum_{i=1}^{4} \log p_{i,y_i}, \tag{3.2}$$

where p_{i,y_i} denotes the predicted probability of the true class y_i for sample i.

Self-Supervised Masked Autoencoder

The model follows the same encoder backbone as the supervised classifier, but it is a self-supervised architecture, as shown in Figure 3.5b. During training, a random binary mask is applied independently to each input window: each time step is set to zero with probability $p_{\text{mask}} = 0.3$. The encoder processes the masked signal and produces a latent vector of dimension $d = 8$ applying global average pooling to the GRU output. A decoder is used to reconstruct the original signal from the latent vector. It consists of a dense layer that expands the latent vector to $T \times 32$ dimensions, followed by a reshape. Two transposed convolutional layers (kernel sizes 3 and 5 respectively, with stride 1) are then used to progressively reconstruct the sequence, followed by a final 6-channel convolutional layer that generates the reconstructed output.

The model is trained to minimize the Mean Square Error (MSE) between the original and reconstructed signals:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{NTC}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\left(x_{n,t,c} - \hat{x}_{n,t,c}\right)^2, \tag{3.3}$$

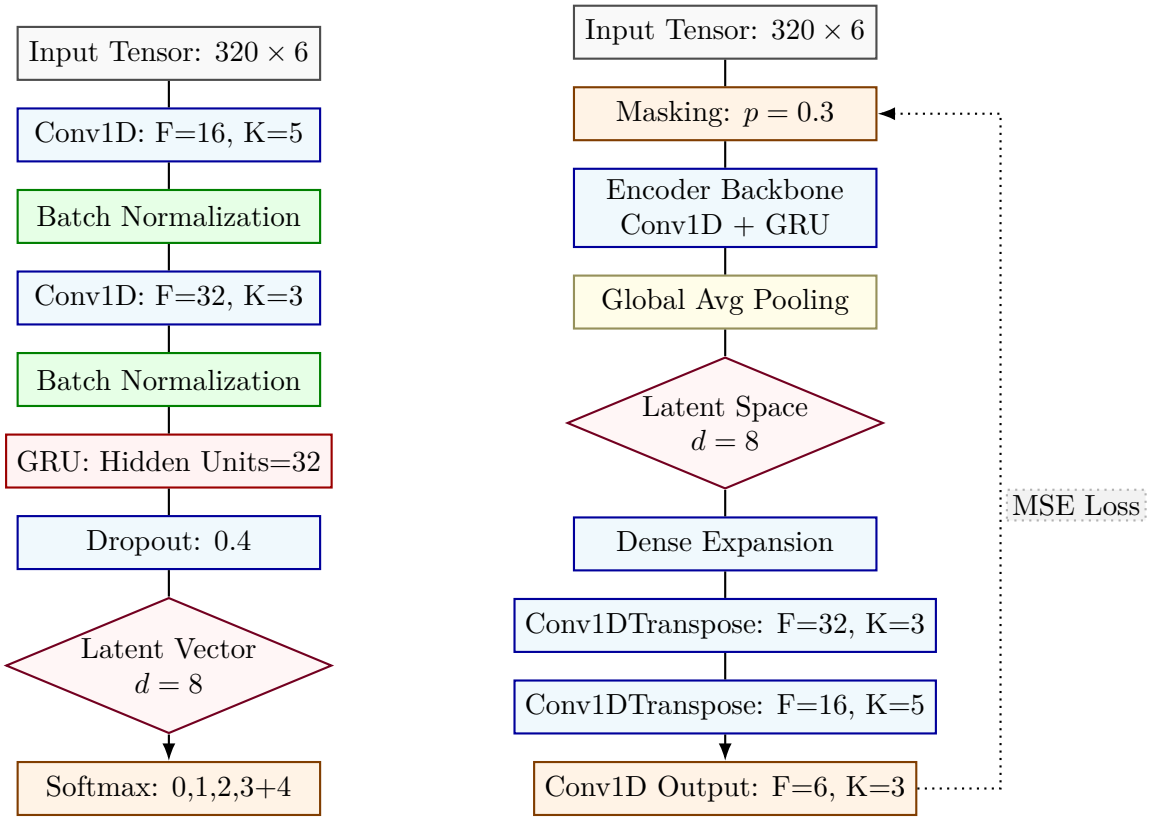
where N is the batch size, T is the sequence length, C is the number of channels, and $x_{n,t,c}$ and $\hat{x}_{n,t,c}$ denote the original and reconstructed signal values, respectively.

Masking is applied only during training, while at inference time the encoder processes the original unmasked signal.

Patient-Level Latent Aggregation

After training, the encoder is used in inference mode to extract the 8-dimensional latent vector from every window. The set of latent vectors corresponding to all windows of a patient is then aggregated into a fixed-size representation by computing four statistics per dimension: mean, standard deviation, maximum and ordinary least-squares linear slope. The slope is computed as:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{N}(i - \bar{i})\, z_{ij}}{\sum_{i=1}^{N}(i - \bar{i})^2}, \tag{3.4}$$



(a) Supervised classifier architecture.

(b) Autoencoder architecture.

Figure 3.5: Encoder architectures for latent feature extraction. (a) Supervised classifier trained end-to-end to predict the four-class MDS-UPDRS Item 3.12 score. (b) Self-supervised masked autoencoder trained to reconstruct randomly masked input segments (masking probability $p = 0.3$) via MSE loss.

where z_{ij} is the j-th latent dimension of the i-th window and $\bar{i}$ is the mean window index. This produces a $4 \times 8 = 32$-dimensional latent feature vector per patient.

3.3.2 Handcrafted Features

Clinically relevant features are extracted independently for stance and walking windows [49]. They are then aggregated at the patient level by computing the mean and standard deviation across all windows of the same task type.

Stance features. Four metrics are computed from each stance window using the medio-lateral (y) and antero-posterior (z) accelerometer channels, after applying a band-pass filter between 0.03–1.0 Hz to isolate postural sway dynamics [23, 49].

The kinematic sway area is estimated as the area of the 95% confidence ellipse fitted to the two-dimensional filtered acceleration signal [9, 23]:

$$A_{\text{sway}} = \pi \cdot 5.991 \cdot \sqrt{\max(\det(\Sigma_{\text{AP,ML}}), 0)}, \tag{3.5}$$

where $\Sigma_{\text{AP,ML}}$ is the 2×2 covariance matrix computed between the filtered antero-posterior and medio-lateral acceleration vectors.

Lateral dominance tracks directionality and is defined as the ratio of the medio-lateral variance to the antero-posterior variance [49]:

$$\text{Lat}_{\text{dom}} = \frac{\sigma^2_{\text{ML}}}{\sigma^2_{\text{AP}} + \epsilon}, \tag{3.6}$$

where $\epsilon = 10^{-8}$ is a small regularization factor to prevent division by zero.

Two spectral power ratios are derived from the discrete fast Fourier transform spectrum of the medio-lateral channel [9, 49]: P_{sway}, representing the fraction of total power falling within the postural sway band (0.1–0.5 Hz), and P_{tremor}, representing the fraction of total power in the tremor band (8–12 Hz). Both are normalized by the total spectral energy.

Aggregating the mean and standard deviation over all stance windows results in $4 \times 2 = 8$ stance features per patient.

Walking features. Three metrics are computed from each walking window [23, 49]. The jerk magnitude uses all three accelerometer channels, while step cadence variability focuses on the vertical (x) and the dominant frequency on the antero-posterior (z) component.

The accumulated jerk magnitude is computed from the discrete first derivative of the triaxial acceleration vector and normalized by the total window duration

[9, 23]:

$$J_{\text{norm}} = \frac{1}{\frac{T}{f_s} + \epsilon} \sum_{t=1}^{T-1} \|(\mathbf{a}_{t+1} - \mathbf{a}_t) \cdot f_s\|_2 \,, \tag{3.7}$$

where $\mathbf{a}_t \in \mathbb{R}^3$ represents the triaxial acceleration sample at time step t, T is the total number of samples in the window, and f_s is the sampling frequency.

Step cadence variability is evaluated as the Coefficient of Variation (CV) of the inter-peak intervals extracted from the vertical acceleration (x) [49]. Peaks are isolated using a minimum distance constraint of $\lfloor 0.3 \cdot f_s \rfloor$ samples and a prominence threshold of $0.3 \cdot \sigma_{\text{vertical}}$. If more than one peak is detected, the feature is calculated as:

$$\text{Step}_{\text{CV}} = \frac{\sigma(\mathbf{\Delta t}_{\text{peaks}})}{\mu(\mathbf{\Delta t}_{\text{peaks}}) + \epsilon}, \tag{3.8}$$

where $\mathbf{\Delta t}_{\text{peaks}}$ is the vector of temporal distances between consecutive peaks expressed in seconds. If one or no peaks are found, Step_{CV} defaults to 0.

Finally, the dominant frequency (f_{dom}) is identified by extracting the frequency corresponding to the maximum magnitude within the physiological band (0.5–4.0 Hz) of the antero-posterior acceleration magnitude spectrum [9, 49].

Aggregating the mean and standard deviation results in $3 \times 2 = 6$ walking features per patient.

The full patient-level feature vector is obtained by concatenating the latent block (32 dimensions), the stance block (8 dimensions), and the walking block (6 dimensions), for a total of 46 dimensions.

3.4 Classification Model

The final patient-level classification task consists of binary prediction of postural instability: stable (MDS-UPDRS Item 3.12 score 0) versus unstable (score ≥ 1). The 46-dimensional patient-level feature vector described in Section 3.3 is fed into four classical ML classifiers. To quantify domain shift, a separate domain classifier is trained on the same features to predict the dataset of origin. Additionally, two adaptation methods are implemented to counter this problem and increase cross-dataset generalization.

Classifiers. Four ML classifiers are compared. First, a RF [12] was employed, with 50 trees, maximum depth 3, and minimum 3 samples per leaf. Second, a GBM [21] composed of 50 shallow trees was trained. Then, an SVM [19] model was used, with Radial Basis Function (RBF) kernel and calibration to obtain soft outputs for ROC-AUC computation. Finally, LR [16] was chosen as standard for binary classification. In all models, balanced class weights are applied to overcome

class imbalance of stable and unstable patients across datasets. The models' hyperparameters were not defined through a structured grid search, most were kept at their default values. For the tree-based methods, a small number of shallow trees and a low minimum number of samples per leaf were chosen to avoid overfitting, given the limited number of features and subjects. The complete set of parameters is reported in Table A.1.

Feature Subsets and Selection. Four feature subsets are evaluated for each classifier. First, the 8 stance and 6 walking descriptors described in Subsection 3.3.2, are fed separately to the classification models to evaluate their discriminative power. Then, following the same reasoning, the classifiers are trained only using the aggregated latent statistics extracted from the encoder, presented in Subsection 3.3.1. Finally, the concatenated vector composed both of latent and handcrafted features is leveraged. Moreover, because the combined feature vector has 46 dimensions and the training contains only 88 subjects, Recursive Feature Elimination (RFE) is applied to reduce the feature space and the risk of overfitting. Its main parameters are reported in Table A.1.

Before training the classifier, zero-variance features are removed using a variance threshold filter ($\sigma^2 < 10^{-6}$), and all remaining features are standardized to zero mean and unit variance using a scaler fitted on the concatenated training and validation sets and then applied to the test set.

3.5 Domain Adaptation

Features extracted from different datasets can retain domain-specific properties due to different hardware, participant demographics, and recording protocols, a problem known as domain shift. A domain classifier was implemented to measure this effect in the extracted features: the same 46-dimensional patient-level feature vectors described in Section 3.3 are used as input, but the target label is replaced with the source dataset (WearGait-PD, Kiel, or OmniaPark) instead of the postural instability score. If the classifier achieves high accuracy, it means the features still carry strong dataset-specific information.

Additionally, to reduce this domain shift at the feature level, two distribution alignment methods are applied to the 46-dimensional patient feature vectors after extraction. In both cases, WearGait-PD is used as the target domain, as it is the largest and most clinically comprehensive dataset in the collection. All alignment parameters are estimated exclusively on the training set and then applied to the test set, to avoid data leakage.

Correlation Alignment (CORAL) [52] aligns the second-order statistics of each source domain s to those of the target domain t. The transformation is:

$$\tilde{x}_s = (x_s - \mu_s)\, C_s^{-1/2}\, C_t^{1/2} + \mu_t, \tag{3.9}$$

where μ_s and μ_t are the source and target feature means, C_s and C_t are the corresponding regularized covariance matrices. This transform whitens the source distribution and then re-colors it to match the target covariance, while the target domain is left unchanged.

Maximum Mean Discrepancy (MMD) alignment [27] is implemented here as a first-order mean-shift. For each source domain s, a constant shift vector is computed as the difference between the target and source feature means:

$$\Delta_s = \mu_t - \mu_s, \tag{3.10}$$

and applied additively to all source feature vectors: $\tilde{x}_s = x_s + \Delta_s$. This approach aligns only the first moment of each source distribution to the target, and is more robust than CORAL when the number of available training samples per domain is small.

3.6 Training and Evaluation

Each binary classifier is trained on the concatenated training and validation sets used by the encoders and evaluated on the held-out test set. The models are evaluated in an ablation study that compares encoder architecture (supervised classifier or self-supervised autoencoder), feature subset (latent only, handcrafted only or combined), domain adaptation strategy (baseline, CORAL or MMD) and finally the ML classifiers described above.

Additionally, each adaptation variant (baseline, CORAL, and MMD) is evaluated independently using the dataset classifier, to directly compare their effect. The experiment is repeated for each combination of encoder architecture (supervised classifier and autoencoder), feature block (Latent, Handcrafted, and Combined), and domain adaptation variant (baseline, CORAL, and MMD). The same four classifiers used in the main task are employed: RF, GBM, SVM, and LR.

The following metrics are adopted to provide a comprehensive evaluation.

Precision and Recall. Recall (true positive rate) measures the fraction of actual unstable patients correctly identified, while precision (positive predictive value) measures the fraction of predicted unstable patients who are actually unstable:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.11}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. In the clinical context of postural instability assessment, high recall is especially important to avoid missing patients who require intervention.

Balanced Accuracy. It represents the arithmetic mean of sensitivity and specificity:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right), \tag{3.12}$$

Unlike standard accuracy, balanced accuracy assigns equal weight to each class regardless of its prevalence [47].

Macro F1 Score. The unweighted mean of the per-class F1 scores:

$$F_1^{\text{macro}} = \frac{1}{2}\sum_{c\in\{0,1\}} \frac{2\cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{3.13}$$

where precision and recall are computed independently for each class and then averaged. Macro F1 penalizes classifiers that achieve high performance on only one class [47].

ROC-AUC. It is computed from the classifier's soft probability output. ROC-AUC does not depend on the threshold and measures the classifier's ability to rank positive instances above negative ones across all possible decision thresholds. A value of 0.5 corresponds to random performance, while 1.0 means perfect classification [47].

Precision-Recall Curve. It is computed from the predicted class probabilities by plotting precision against recall across all possible decision thresholds. Unlike ROC-AUC, it focuses on the positive class and is particularly informative under class imbalance, where it better reflects performance on the minority class. The baseline corresponds to the positive-class prevalence, while a value of 1.0 means perfect classification.

To obtain reliable uncertainty estimates, a bootstrap procedure with $N = 1{,}000$ iterations was applied to the fixed seed-42 test set. Each iteration substitutes a single randomly chosen sample with another drawn uniformly from the test set. All Confidence Intervals (CIs) are reported at the 95% level (2.5^{th}–97.5^{th} percentiles of the bootstrap distribution).

3.7 Ordinal Severity Estimation

The main classification task of this study is binary (Stable vs. Unstable). However, it can be interesting to assess whether the continuous probability scores produced by the binary models carry implicit ordinal information about the MDS-UPDRS Item 3.12 scale (0 = Stable, 1 = Mild, 2 = Moderate, 3–4 = Severe). A simple exploratory analysis was conducted to address this question. The predicted probability of instability $p \in [0,1]$ from the best model configuration was treated as an ordinal score. Three thresholds $t_1 < t_2 < t_3$ were defined to derive the four classes from score p:

$$\hat{y} = \begin{cases} 0 & \text{if } p < t_1 \\ 1 & \text{if } t_1 \leq p < t_2 \\ 2 & \text{if } t_2 \leq p < t_3 \\ 3 & \text{if } p \geq t_3. \end{cases} \tag{3.14}$$

The thresholds were defined through a grid search using 40 values evenly spaced over the interval $[0.05, 0.95]$. All configurations satisfying $t_1 < t_2 < t_3$ were tested and the corresponding results were evaluated using the balanced accuracy against the true MDS-UPDRS Item 3.12 values. Balanced accuracy, F1 score, confusion matrix and probability distributions are reported for the best set of thresholds. It should be noted that the thresholds are optimized directly on the test set and no actual training or validation is performed.

Chapter 4

Results

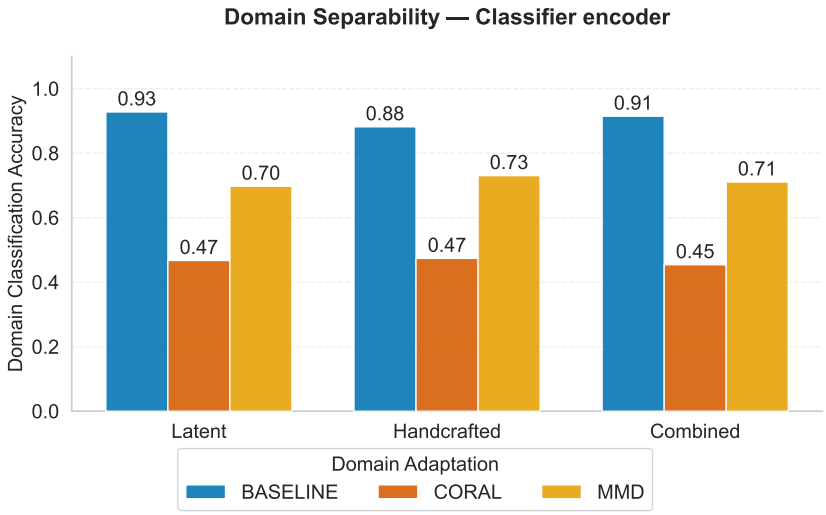
All results reported in this chapter are obtained on a fixed random seed (42), used to ensure reproducibility. Specifically, the seed controls the train/validation/test split, the data shuffling, the neural network training, and the initialisation of all ML classifiers. The test set consists of 38 patients, held out before any training or feature extraction. Given the limited number of test examples, bootstrap confidence intervals (95%, $N = 1000$ resamples) are reported for some configurations to quantify estimation uncertainty.

4.1 Domain Shift Analysis

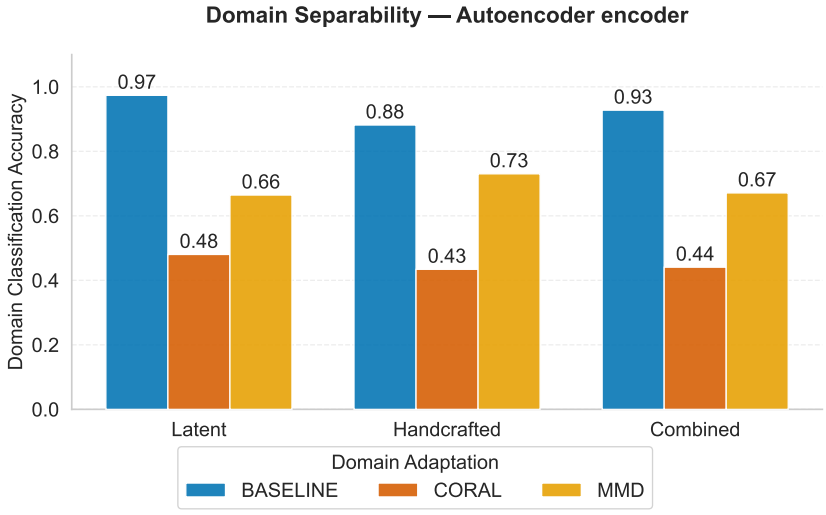
As explained in 3.5, a specific domain classifier is used to evaluate domain shift. Figure 4.1 shows the resulting accuracy for the two encoder architectures, across all feature blocks and domain adaptation variants. The reported values correspond to the mean accuracy over the four classifiers. Without any domain adaptation (Baseline), the domain classifier achieves very high accuracy across all feature blocks and both encoders, ranging from 0.88 to 0.97. CORAL is the most effective method at reducing domain separability, providing a reduction in accuracy in the range 0.43–0.48, which is close to random guessing (33%). MMD also reduces domain information, but less aggressively, with accuracy in the range of 0.66–0.73. This pattern is consistent across both encoder architectures and all feature blocks.

The Uniform Manifold Approximation and Projection (UMAP) projections of the 32-dimensional latent block, shown in Figure A.1, offer a qualitative view of the same effect measured by the domain classifier. In the baseline setting both encoders show dataset clustering, which is most evident for the autoencoder, where the three datasets form clearly separated groups, consistent with the high domain-classifier accuracy of 0.88–0.97. After CORAL alignment the clusters become less evident in both encoders, in agreement with the drop in domain-classifier accuracy

to 0.43–0.48, although some residual structure remains. MMD produces the most uniform distribution for the autoencoder, while for the supervised encoder CORAL and MMD yield similarly mixed projections.



(a) Supervised classifier encoder.



(b) Masked autoencoder encoder.

Figure 4.1: Domain classifier accuracy for each feature block and domain adaptation variant, using features extracted either from the classifier encoder (a) or the autoencoder (b).

4.2 Classification Performance

Tables A.2 and A.3 report the complete set of evaluation metrics for all configurations of the supervised classifier encoder and the autoencoder encoder, respectively. Each row corresponds to one combination of feature subset, domain adaptation variant, and classifier. Configurations are sorted by ROC-AUC in descending order within each domain adaptation block. The best value in each column is highlighted in bold. To summarize the results, Figures 4.2 and 4.3 show the ROC-AUC metric for all combinations of classifier, feature block, and domain adaptation variant, for the classifier encoder and autoencoder respectively.

Effect of the Encoder. The two encoders achieve competitive results, as shown in Figures 4.2 and 4.3. The supervised encoder reaches its best performance using latent features with CORAL adaptation and LR as classifier (ROC-AUC = 0.972). The autoencoder achieves its best result using the combined feature set with RFE and CORAL adaptation, with RF as classifier (ROC-AUC = 0.949). Notice that the classifier encoder achieves generally higher performance, but its margin over the autoencoder is small, even though the latter is trained without any labels.

Effect of the Feature Block. Figures 4.2 and 4.3 show that handcrafted features alone produce the lowest performance in all configurations, with ROC-AUC in the range of 0.57–0.74. Latent features provide a large improvement and the best results is obtained by the classifier encoder combined with CORAL, reaching 0.972. Combined latent and handcrafted features is beneficial for the final classification performance, with equal or slightly better results than latent features only. Moreover, the addition of RFE (Combined_RFE) further improves the results in several configurations, achieving a maximum ROC-AUC of 0.960 (SVM with MMD) for the supervised encoder and 0.949 (RF with CORAL) for the autoencoder. Table 4.1 reports the features retained by RFE across the six encoder/adaptation configurations. Latent dimensions are consistently the most represented, while handcrafted descriptors are chosen differently by the adaptation strategies. In the Baseline setting only two handcrafted features survive (mean postural-sway power ratio and mean lateral dominance for the autoencoder; gait-rhythm descriptors for the classifier encoder). When using CORAL, handcrafted features become more relevant and sway-area, spectral-power, and gait-variability descriptors represent half of the selected subset, while under MMD, only a few stance descriptors are kept.

Effect of the Classifier. No single classifier dominates across all settings, and the differences between them are generally small, as better described in Tables A.2 and A.3. On handcrafted features all four models perform poorly and close to one

Encoder	Adaptation	Handcrafted	# Latent
Autoencoder	Baseline	sway power ratio mean, lateral dominance mean	13
	CORAL	sway area mean, sway area std, lateral dominance mean, sway power ratio mean, sway power ratio std, tremor power ratio mean, dominant gait freq. std	8
	MMD	sway area mean, sway area std, sway power ratio mean, step cadence CV mean	11
Classifier	Baseline	step cadence CV std, dominant gait freq. mean, dominant gait freq. std	12
	CORAL	sway area std, lateral dominance mean, sway power ratio std, tremor power ratio mean, dominant gait freq. std	10
	MMD	sway area std, sway power ratio std, tremor power ratio mean, step cadence CV std	11

Table 4.1: Handcrafted features selected by RFE across all six configurations (15 features total per configuration).

another (ROC-AUC in the 0.67–0.74 range). On the latent and combined feature blocks, the linear and margin-based classifiers tend to be the most competitive: LR reaches the best result for the supervised encoder (0.972 on latent features) and is consistently among the top models for both encoders, while SVM achieves the highest value with Combined_RFE (0.960). The tree ensembles behave differently across encoders, with RF performing best for the autoencoder (0.949 with Combined_RFE) but worse for the supervised one, and GBM remains competitive but never returns the best result.

Effect of Domain Adaptation. Comparing Figures 4.1 with 4.2 and 4.3, it can be noticed that the effect of domain adaptation on classification performance is different from its effect on domain shift. CORAL is the most effective at removing domain information, but this does not always translate into better classification performance. In some configurations it achieves the highest ROC-AUC: for example, the supervised encoder with latent features, CORAL adaptation, and LR as classifier reaches 0.972. In others, however, it performs similarly to or worse than the baseline: for instance, using latent features with CORAL and SVM, the ROC-AUC drops to 0.776, compared to 0.849 without adaptation. MMD shows a more consistent improvement over the baseline across configurations, particularly for the autoencoder encoder.

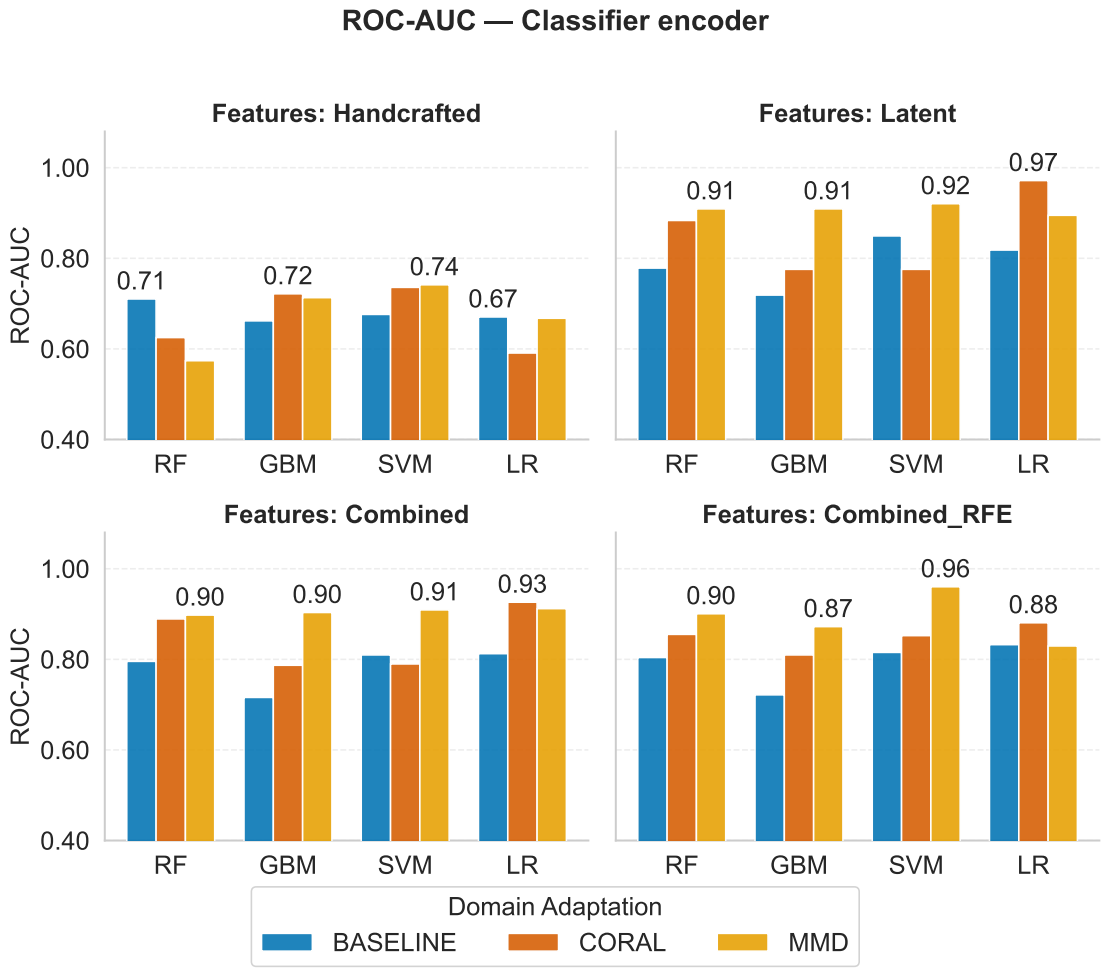


Figure 4.2: ROC-AUC across classifiers (RF, GBM, SVM, LR) and feature blocks for each domain adaptation variant, supervised classifier encoder (seed 42).

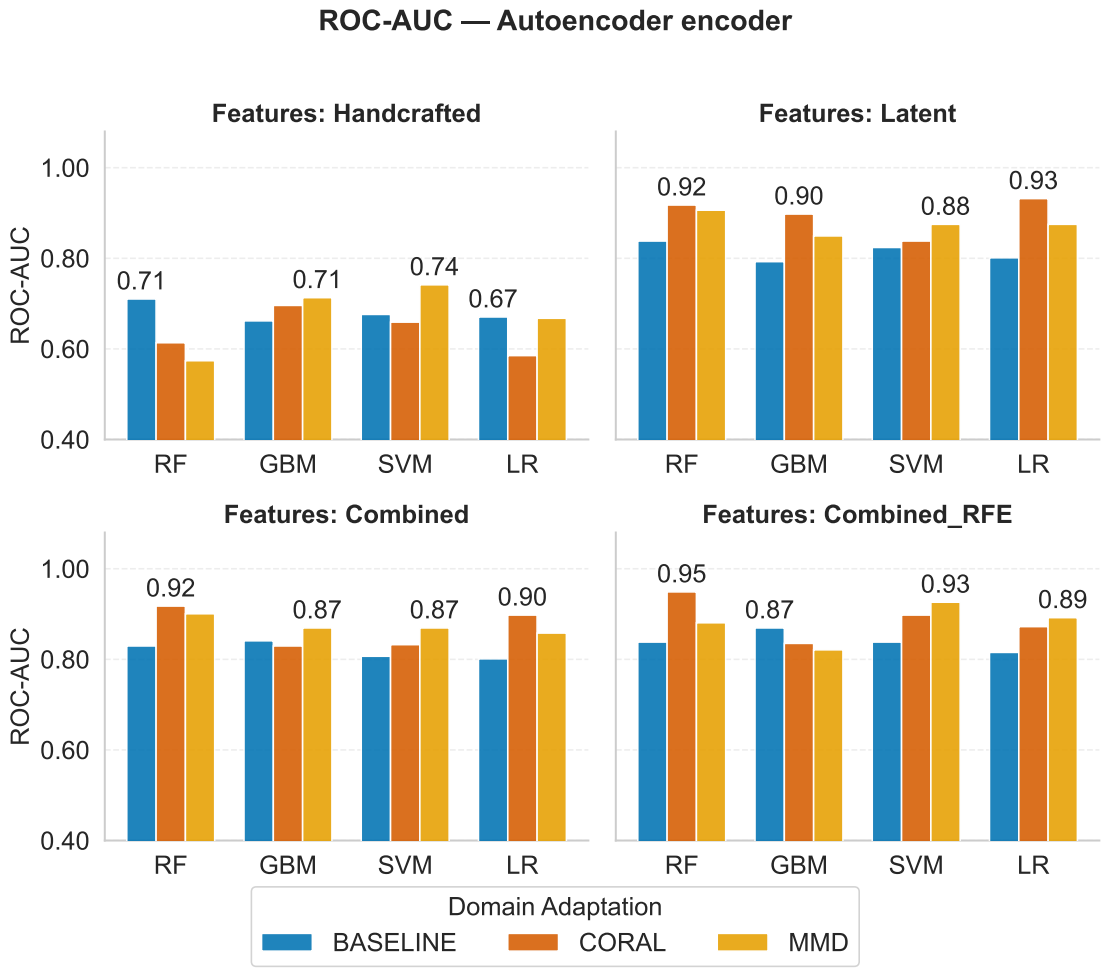


Figure 4.3: ROC-AUC across classifiers (RF, GBM, SVM, LR) and feature blocks for each domain adaptation variant, autoencoder encoder (seed 42).

4.2.1 Bootstrap Confidence Intervals

As explained in Section 3.6, CIs were computed to obtain reliable uncertainty estimations. The results for the two best configurations (classifier encoder, latent features, LR classifier and CORAL, and autoencoder encoder, combined features with RFE, RF classifier and CORAL) are shown below.

ROC Curves. Figure 4.4 reports the mean ROC-AUC curve and its 95% bootstrap confidence band for each configuration. The classifier encoder, with latent features, LR and CORAL achieves a ROC-AUC of 0.972 (95 % CI [0.962, 0.988]), while the autoencoder, with Combined-RFE, RF and CORAL yields a ROC-AUC of 0.949 (95 % CI [0.930, 0.966]). Both curves indicate high true-positive rates at low false-positive rates, and their confidence bands are quite narrow.

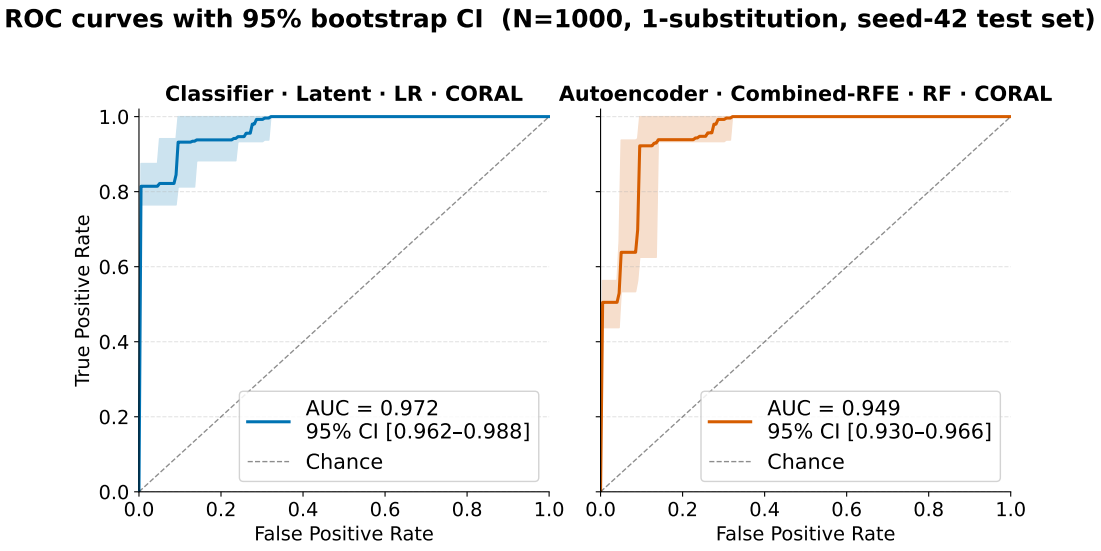


Figure 4.4: ROC curves with 95 % bootstrap confidence bands for the two selected configurations ($N = 1{,}000$ iterations, 1-substitution, seed-42 test set).

Precision-Recall Curves. Figure 4.5 shows the corresponding Precision-Recall curves. The positive-class prevalence in the test set is approximately 0.42, represented by the horizontal dashed chance baseline. The classifier encoder, with Latent features, LR, CORAL configuration obtains an Average Precision of 0.968 (95 % CI [0.955, 0.984]). On the other hand, the autoencoder, with Combined-RFE, RF and CORAL configuration achieves 0.930 Average Precision (95 % CI [0.902, 0.957]).

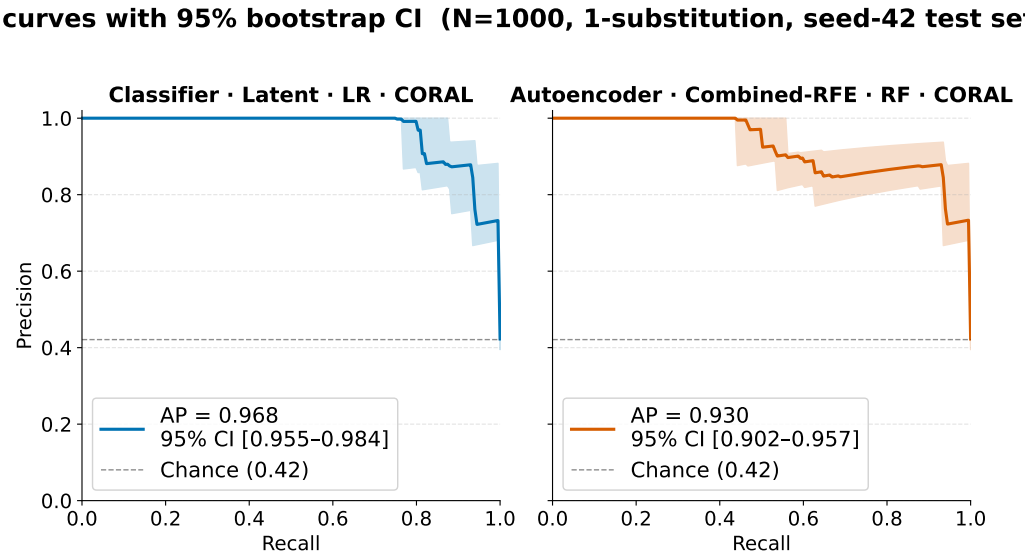


Figure 4.5: Precision-Recall curves with 95 % bootstrap confidence bands for the two selected configurations ($N = 1{,}000$ iterations, 1-substitution, seed-42 test set).

4.2.2 Confusion Matrices and Probability Distributions

Figures 4.6, 4.7, 4.8, and 4.9 show the confusion matrices and predicted probability density distributions for the top-4 configurations under CORAL adaptation, for the supervised encoder and the autoencoder respectively. These configurations are selected as the most informative given their highest overall ROC-AUC.

The confusion matrices show that the best configurations produce very few misclassified examples. For the best supervised encoder configuration (latent features, CORAL adaptation, and LR as classifier) only 2 stable patients are misclassified as unstable and 1 unstable patient is misclassified as stable, out of 38 test subjects.

The probability density plots show the predicted probability of instability, colored by the original MDS-UPDRS score (0, 1, 2, 3–4). In the best configurations, the stable class (score 0) is well separated to the left of the 0.5 decision threshold, while severe cases (score 3–4) concentrate strongly to the right. The intermediate classes (scores 1 and 2) overlap with each other and partially with the stable class, which is clinically expected given their borderline nature.

The complete set of metrics for the two best configurations is reported in Table 4.2.

Metric	**Classifier, Latent, LR, CORAL**	**Autoencoder, Combined_RFE, RF, CORAL**
ROC-AUC	0.972	0.949
AUC-PR	0.968	0.930
Bal. Acc.	0.923	0.901
Macro F1	0.920	0.894
Precision	0.882	0.833
Recall	0.938	0.938

Table 4.2: Complete set of metrics for the two best configurations (classifier encoder, latent features, LR and CORAL, and autoencoder encoder, combined features with RFE, RF and CORAL)
Abbreviations: RFE = Recursive Feature Elimination; RF = Random Forest; LR = Logistic Regression; ROC-AUC = Receiver Operating Characteristic Area Under the Curve; AUC-PR = Area Under the Precision-Recall Curve; Bal. Acc. = Balanced Accuracy; Macro F1 = macro-averaged F1 score.

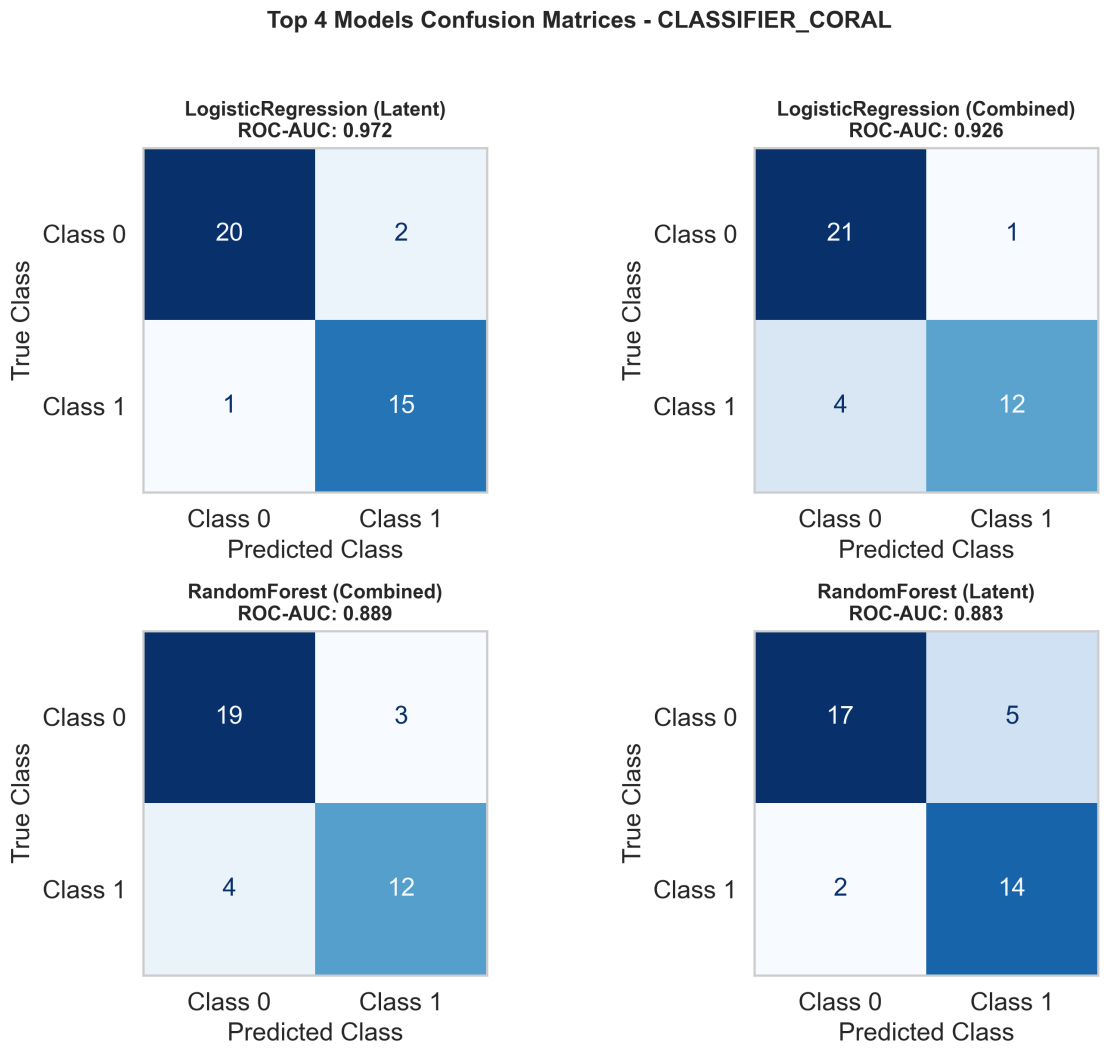


Figure 4.6: Confusion matrices for the top-4 model configurations under CORAL adaptation — supervised classifier encoder.

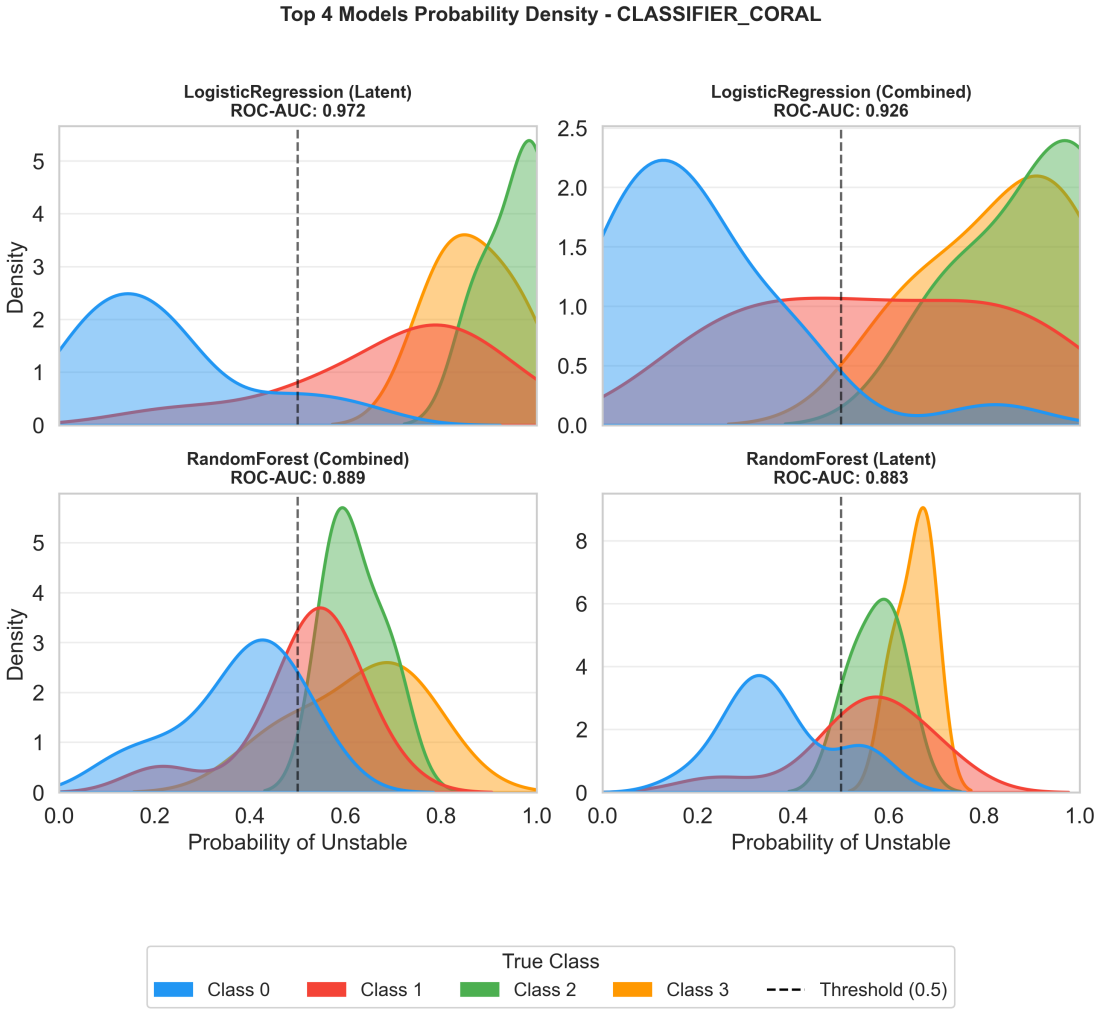


Figure 4.7: Predicted probability density of the predicted instability score for the top-4 model configurations under CORAL adaptation — supervised classifier encoder.

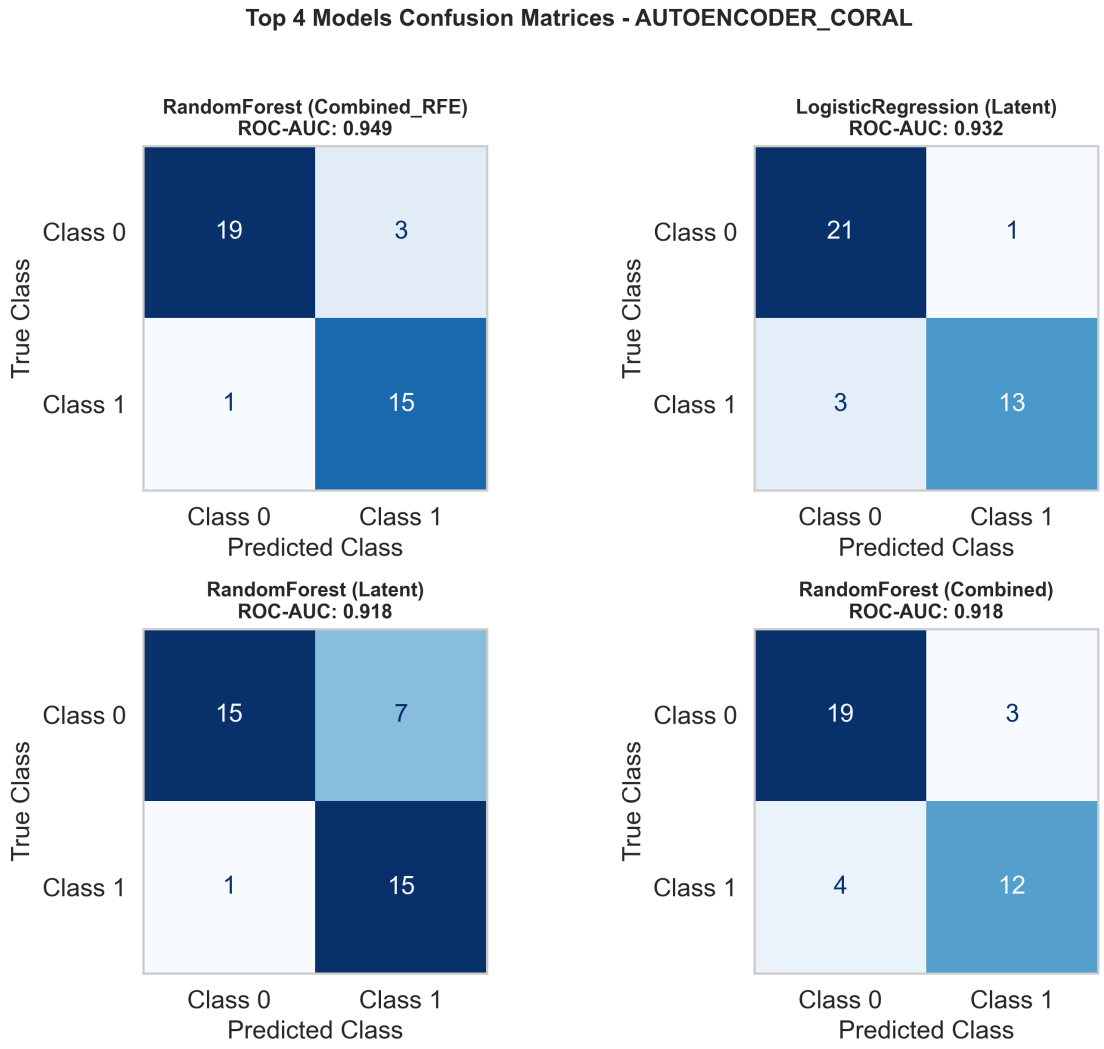


Figure 4.8: Confusion matrices for the top-4 model configurations under CORAL adaptation — autoencoder encoder.

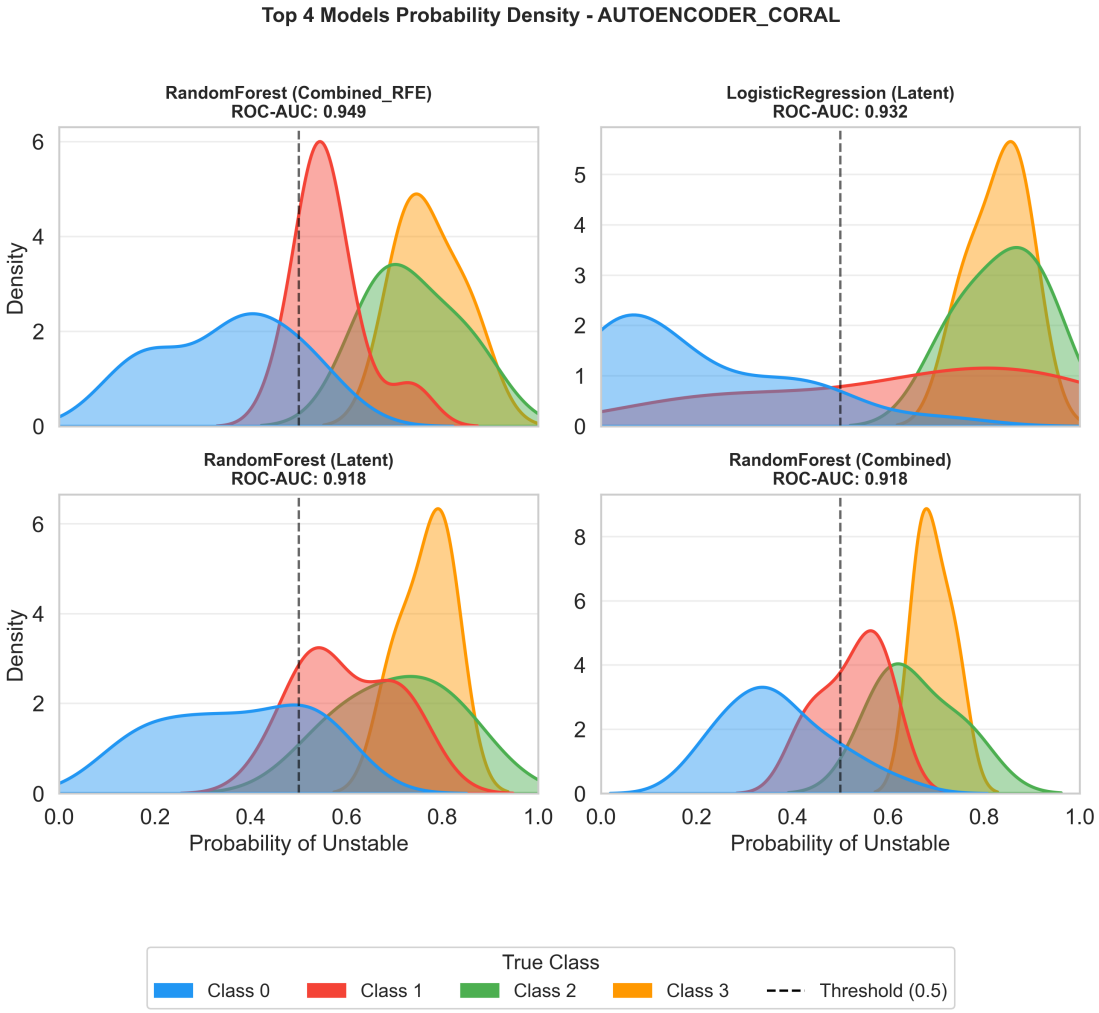


Figure 4.9: Predicted probability density of the predicted instability score for the top-4 model configurations under CORAL adaptation — autoencoder encoder.

4.3 Ordinal Severity Estimation

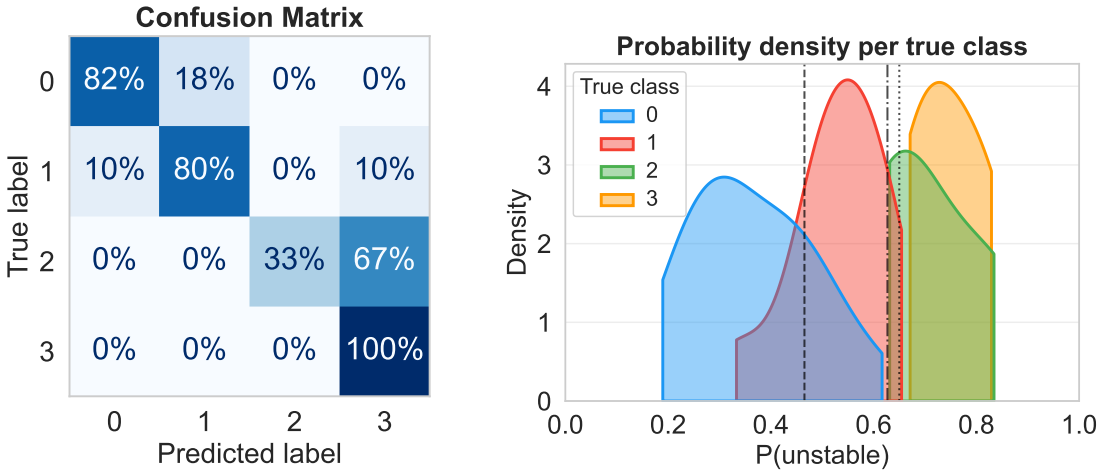
The ordinal severity analysis described in Section 3.7 was applied to the probabilities obtained by the autoencoder architecture, RF model with combined features selected through RFE and CORAL adaptation. This configuration was chosen because its probability outputs already appear well calibrated and preserve the correct ordering of the severity classes, as shown in Figure 4.9.

The thresholds obtained using the grid search approach are as follows:

$$\hat{y} = \begin{cases} 0 & \text{if } p < t_1 \\ 1 & \text{if } t_1 \leq p < t_2 \\ 2 & \text{if } t_2 \leq p < t_3 \\ 3 & \text{if } p \geq t_3 \end{cases} \quad t_1 = 0.51,\ t_2 = 0.63,\ t_3 = 0.67. \tag{4.1}$$

Figure 4.10 shows the probability distribution per true severity class and the normalized confusion matrix. The four classes separate along the probability axis, with mean scores of 0.35 ± 0.14 (Stable), 0.56 ± 0.07 (Mild), 0.74 ± 0.08 (Moderate), and 0.77 ± 0.06 (Severe). This ordering suggests that the learned feature implicitly carry disease severity information.

The four-class classification achieves a balanced accuracy of 0.761 and a Macro F1 of 0.724. Stable (Class 0) and Severe (Class 3) patients are classified with high recall (91% and 100% respectively), reflecting the strong separation of the two extremes of the severity spectrum. Mild patients (Class 1) achieve 80% recall, with 10% misclassified as Stable and 10% as Severe. Moderate patients (Class 2) show 33% recall, with 67% misclassified as Severe. However, this result is statistically unreliable given the very small sample size ($n = 3$).



(a) Confusion matrix.

(b) Probability density per true class.

Figure 4.10: Exploratory ordinal severity estimation from binary $P(\text{Unstable})$ scores.

4.4 Comparison with Similar Studies

Table 4.3 compares this work with the most relevant prior studies on IMU-based postural stability assessment. It should be stressed that performances alone should only be considered indicative since the studies target different clinical scales (MDS-UPDRS Item 3.12, H&Y stage, or BBS), adopt different acquisition protocols, and, most importantly, evaluate within a single dataset rather than across datasets.

One of the most relevant differences concerns the data. All prior works rely on a single dataset acquired at a single clinical site, with cohorts of 30–42 PD patients [9, 49]. In contrast, this work merges three heterogeneous datasets (WearGait-PD [1], Kiel [56], and OmniaPark [22]) collected at multiple sites across the United States, Germany, and Italy, for a total of 149 PD patients. To the best of our knowledge, this is the first study to address postural instability assessment in PD in a cross-dataset setting and to quantify and address the domain shift.

Moreover, some fundamental methodological differences from previous works emerge. The reviewed studies can be divided into two main groups: pipelines based exclusively on handcrafted descriptors fed to a classical model or statistical test [9, 49], and fully end-to-end deep architectures [33]. In contrast, this work adopts a hybrid strategy in which a deep encoder is used as a feature extractor and its latent vectors are aggregated at the patient level and used as input features for a classical ML classifier. A natural design question is whether this hybrid approach is really needed, or a fully DL classification pipeline would achieve comparable performances. The training history of the supervised CNN-GRU encoder provides a direct answer. Figure 4.11 shows that the window-level four-class classifier reaches only ∼39% training accuracy and ∼35% validation accuracy at convergence.

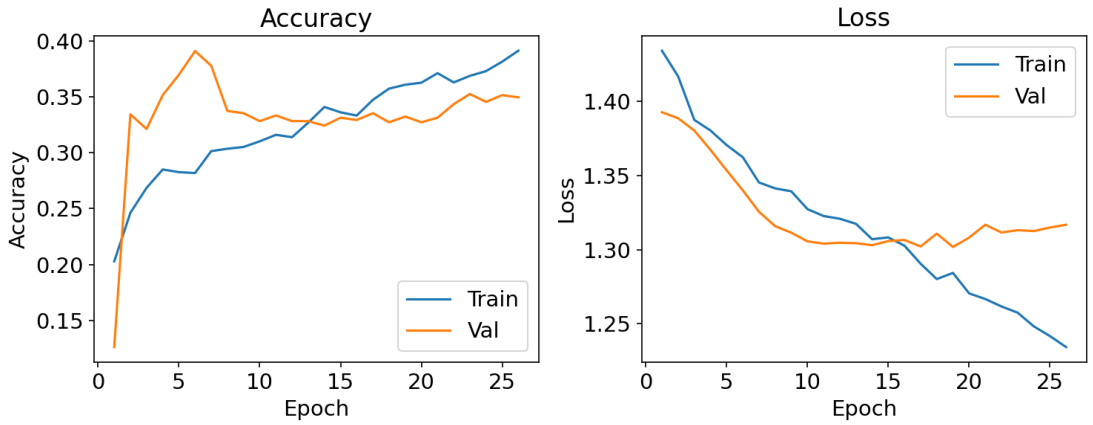


Figure 4.11: Training and validation accuracy (a) and loss (b) of the supervised CNN-GRU window-level classifier.

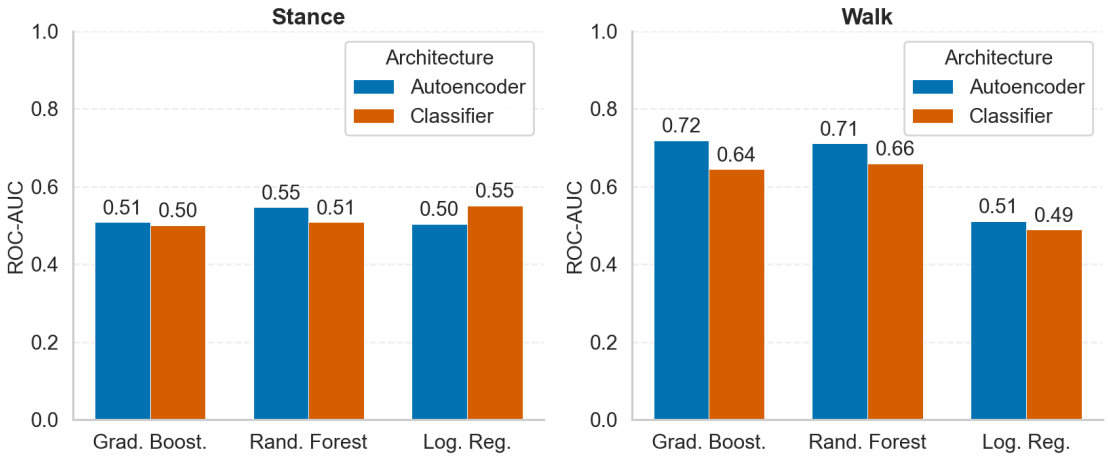


Figure 4.12: ROC-AUC obtained when the pipeline is applied separately to stance-only (a) and walk-only (b) windows, using the latent feature block without domain adaptation. Results are shown for three classifiers (Gradient Boosting, Random Forest, Logistic Regression) and both encoder architectures (Autoencoder and Classifier).

A further methodological difference concerns the task choices. Prior studies targeting postural instability in PD rely on a single task, either quiet stance in Borzì et al. [9] and in the descriptors surveyed by Ghislieri et al. [23] or the pull test in Smith et al. [49]. In this work instead both quiet stance and walk are leveraged. To further motivate the use of both stance and walk tasks, Figure 4.12 reports the ROC-AUC obtained when the same pipeline is applied separately to stance-only and walk-only windows, using the latent feature block without domain adaptation. Stance-only performance is quite low across all classifiers and both encoder architectures, reaching at most 0.55 (Random Forest, Autoencoder). Walk-only results are higher but still limited, with a maximum of 0.72 (Gradient Boosting, Autoencoder), compared to a mean ROC-AUC of 0.79–0.84 for the joint Combined configuration at baseline.

Additionally, the test and validation strategy used in this work differs from previous studies. Borzì et al. [9] adopt a Leave-One-Subject-Out (LOSO) approach but on a single dataset and perform no external validation. Smith et al. [49] report discriminative power through statistical feature ranking on 30 patients and explicitly state that multi-site validation on larger cohorts is still required. Kim et al. [33] use 10-fold cross-validation but at the task/window level on a non-PD population. This work instead relies on a patient-level hold-out strategy, with a test set of 38 PD patients that is isolated before any training, feature extraction, or feature selection. This test set is substantial for this type of study: it is comparable in size to the entire cohorts on which the most relevant prior works are developed and evaluated (42 and 30 PD patients, respectively [9, 49]). Estimation uncertainty on this held-out set is further characterized through 1,000-iteration bootstrap confidence intervals.

Finally, the obtained results can be compared with the ones of previous works. The cross-dataset ROC-AUC of 0.972 (95% CI [0.962, 0.988]) obtained by the best configuration is competitive with the best results reported in the literature, such as the 0.95 ROC-AUC of Smith et al. [49], and exceeds the only comparable multiclass PD result, the 70.1% accuracy of Borzì et al. [9]. Notably, even without any domain adaptation the pipeline already reaches a ROC-AUC of 0.85. The excellent scores reported elsewhere (100% in [9], 98.4% in [33]) are obtained either on a single site or at the task level on a non-PD cohort, and are not directly transferable to the cross-dataset, patient-level problem addressed here.

Reference	Data & cohort	Sensors	Tasks	Approach	Validation & test set	Target & main result
Borzì et al. 2020 [9]	1 dataset, single site (Turin); 42 PD + 7 HC	1 smartphone (acc + gyr), lower back	Quiet stance	414 handcrafted feat. → Pearson selection (8); linear SVM, two-layer cascade	LOSO; *in-dataset*, no external validation	MDS-UPDRS 3.12. HC vs. PS Acc = 100%; multi-class Acc = 70.1%, F1 = 67.3%
Smith et al. 2025 [49]	1 dataset, single site (Madrid); 30 PD (H&Y II vs. III)	6 IMUs (acc); chest–lumbar–shanks–feet	Pull test	60 handcrafted feat.; statistical ranking (Mann–Whitney U + ROC-AUC); no learned classifier	Feature ranking; *in-dataset*; authors note multi-site validation still needed	H&Y II vs. III. Best *single-feature* AUC = 0.95
Kim et al. 2021 [33]	1 dataset, single site (Korea); 53 subjects (stroke rehab + healthy, *not PD*)	1 IMU (acc + gyr), trunk	14 BBS tasks	End-to-end DL ensemble: 2×1D-CNN + GRU + dense meta-learner	10-fold CV; *window/task-level*, in-dataset	BBS item score (0–4). Avg Acc = 98.4% across 14 tasks
This work	**3 datasets, multi-site** (USA, Germany, Italy); **149 PD**	**1 IMU (acc + gyr), lower back**	**Quiet stance + gait**	**Hybrid**: DL encoder (supervised CNN-GRU *or* self-supervised masked AE) → 32 latent + 14 handcrafted feat.; domain adaptation (CORAL / MMD); classical ML	Patient-level **held-out test (38 PD)**; bootstrap 95% CI; explicit domain-shift quantification	MDS-UPDRS 3.12 (binary). **Cross-dataset ROC-AUC = 0.97**; 0.85 without adaptation; 4-class bal. Acc = 0.76

Table 4.3: Comparison of this work with representative prior studies.
Abbreviations: PD = Parkinson's Disease; HC = Healthy Controls; PS = Postural Stability; acc = accelerometer; gyr = gyroscope; H&Y = Hoehn & Yahr; BBS = Berg Balance Scale; MDS-UPDRS = Movement Disorder Society Unified Parkinson's Disease Rating Scale; feat. = features; SVM = Support Vector Machine; LOSO = Leave-One-Subject-Out; DL = Deep Learning; CNN = Convolutional Neural Network; GRU = Gated Recurrent Unit; AE = Autoencoder; CORAL = Correlation Alignment; MMD = Maximum Mean Discrepancy; ML = Machine Learning; CV = Cross-Validation; CI = Confidence Interval; ROC-AUC = Receiver Operating Characteristic Area Under the Curve; Acc = Accuracy; F1 = F1 score; bal. Acc = Balanced Accuracy.

Chapter 5

Discussion

This chapter interprets the findings presented in Chapter 4 and situates them within the existing literature. Notably, the CIs for the two best configurations appear quite narrow, indicating that point estimates are reliable despite the limited test size. This supports the validity of the comparisons discussed below.

Cross-Dataset Generalization and Domain Shift. The primary contribution of this work is that postural instability in PD can be assessed across heterogeneous clinical datasets from a single lower-back IMU. Previous studies developed and validated their models within a single cohort acquired at one site [9, 49]. This approach fails to evaluate the generalization capability of the proposed approach. Cross-center generalization is investigated in this work by analysing domain shift. The results show that without alignment, the originating datasets are almost perfectly recoverable from the features. This can be due to the different sensor characteristics, sample, and experimental protocol. The same difficulty has been reported in the related problem of FOG detection, where a model trained on one dataset transferred poorly to others even after the sampling rate and signal scale had been harmonized [48]. The present results show that once the domain shift is explicitly quantified and corrected, a single-sensor model can obtain performances comparable to state-of-the-art approaches on one dataset. This is also the reason why the two adaptation strategies are discussed separately later in this chapter, since domain separability and classification performances are impacted differently.

Role of the Encoder. A main point of the study is the comparison between a supervised CNN-GRU encoder and a self-supervised masked autoencoder. The two achieve almost equivalent performances, even if only the former is trained using ground-truth labels. This indicates that the inertial signal from the lumbar IMU already encodes the postural information relevant to the task, and that

this information can be recovered through a reconstruction of the signal, without clinical annotation.

This result has a practical relevance because in the clinical field labeled data are scarce and expensive. The finding aligns with previous works, where masked pretraining has been shown to learn useful representations of inertial signals that match or even outperform supervised and contrastive methods [14, 38]. This suggests that a self-supervised encoder could first be pretrained on the large amounts of unlabeled IMU data, which are relatively easy to collect, and then reused for postural instability assessment on the small labeled datasets that are available in clinical practice.

Role of the Feature Representation. The ablation across feature subsets reveals a consistent ordering: handcrafted descriptors achieve the lowest performance, latent features substantially improve on it, and their combination with feature selection proves the best.

Handcrafted features are clinically interpretable and widely used in the balance literature [23, 40], but in a cross-dataset approach they are less effective. Quantities such as sway area or spectral-power ratios depend on the sampling rate, the sensor, and the acquisition protocol, so the same physical behavior yields different values across datasets even after resampling. This also explains why domain adaptation is not very effective on them. The latent representation, by contrast, can capture temporal patterns of sway and gait that are not conveyed by clinical metrics, providing a large improvement. The fact that the concatenation of the two performs best, especially after feature selection, indicates that the clinical and latent representations are partly complementary, and that RFE is able to retain the handcrafted descriptors that are most robust to domain shift rather than discarding all handcrafted information.

Role of the Classifier. No classifier dominates across all configurations, and the gaps between them are generally small. The most consistent pattern is that the lower-capacity linear model (LR) is competitive with, and often superior to, the tree ensembles. This is coherent with the size of the available data: with few subjects complex classifiers are more prone to overfitting and to amplifying domain shift, while a simpler decision rule generalizes better to the unseen target dataset. The fact that SVM and LR achieve the highest performances suggests that the latent space is already close to linearly separable, and that more complex classifiers cannot exploit their additional capacity effectively, given the limited sample size.

Role of Domain Adaptation. Both alignment strategies improve over the baseline, but they behave differently. CORAL is the more aggressive method and removes dataset-specific information effectively. However, this correction does

not necessarily correspond to better classification, and in some configurations it even lowers performance relative to the baseline. This trade-off is predictable: aligning the marginal feature distributions is beneficial only when the class proportions are comparable in all domains, but when they differ, as in this case, it can modify the decision boundary and increase the target error [57]. Consistently, CORAL performs best on the latent features, where the second-order statistics can be estimated more reliably. The MMD mean-shift alignment applies a weaker first-moment correction. Its peak performance is slightly lower, but its effect is more consistent across classifiers and feature subsets. In this sense, MMD appears to be the safer choice when the number of subjects per domain is small and the label distributions are imbalanced, while CORAL is preferable when the goal is to maximize peak performance on a specific, well-characterized configuration. Notably, the confidence intervals for these two best configurations are quite narrow, indicating that the point estimates are reliable despite the limited size of the test set.

Clinical Interpretation of the Predictions. Although the model is trained only to separate stable from unstable patients, its continuous output preserves the ordering of the MDS-UPDRS Item 3.12 scale. Most importantly, classes 0 and 1 are very well discriminated: this is the primary clinical objective, as it corresponds to distinguishing between stable patients and the ones with initial impairment. Notice that also Classes 1 and 2 are clearly separated. Most errors occur between Class 2 and Class 3, which is expected given the subjective nature of the pull-test evaluation. These errors correspond to Class 2 subjects being assigned to Class 3, i.e. an overestimation of severity, which is not conservative. Since scores 3 and 4 were merged because of the scarcity of subjects with the highest severity, the distinction between these two extreme grades cannot be assessed here. Overall, the separation between clearly stable and clearly impaired patients is maintained and misclassifications are concentrated among adjacent categories, suggesting that a dedicated ordinal formulation is a natural extension of this work.

Hybrid Design. A further contribution concerns the rationale for the hybrid design. The deep model trained directly on the MDS-UPDRS labels at the window level achieves poor performance, while aggregating its learned representations at the patient level proves effective. From a clinical point of view, this result is coherent: postural instability is a patient-level trait, and a few seconds of inertial signal during standing or walking do not contain enough information to correctly discriminate impairment levels. The symptom emerges from several trunk dynamics rather than from isolated events [37]. Using the deep network as a feature extractor and then a patient-level classifier to assign a score therefore matches the nature of the target.

Effectiveness of Combined Tasks. A related methodological point concerns the choice of motor tasks. Prior studies rely on a single task, either quiet stance [9, 23] or the pull test [49]. The single-task has shown that in this cross-dataset approach neither stance nor gait alone is sufficient. This suggests that the two tasks capture partly complementary aspects of postural control, consistent with the clinical picture of PD, where balance impairment manifests both as an inability to maintain a stable posture and as instability during locomotion [37]. Moreover, the low stance-only performance may be due to the fact that recordings are shorter and have lower amplitude. This likely makes the extracted features more affected by differences between sensors and protocols in a cross-dataset setting. These results indicate that combining tasks is not only a way to increase the amount of data per patient, but a way to combine different description of their impairment.

Limitations. Despite the promising results, several limitations of this work must be acknowledged. The first concerns the size of the test set. Relative to this specific clinical field, 38 patients is a substantial number, comparable in size to the entire cohorts of the most relevant prior studies. However, it remains small by the standards of ML, where reliable estimates are based on larger samples. This is a structural characteristic of the clinical domain rather than a methodological choice, since collecting labeled IMU data in PD requires expert neurologists, which limits the scale of the available cohorts.

Second, all results are reported for a single fixed split, controlled by one random seed. Although the seed is fixed only to ensure reproducibility, the reported performance may depend on the particular train/test partition it induces. Future work should therefore repeat the evaluation across multiple seeds, or adopt a cross-validation scheme, to confirm that the results are stable and not specific to this split, a strategy that was not feasible here given the limited number of subjects.

Additionally, the ground-truth labels are derived from the MDS-UPDRS Item 3.12 score, an ordinal clinical rating assigned by experts and therefore subject to variability. The ground truth thus inherits a level of label noise, that can only be partly reduced by aggregating information at the patient level. Moreover, intermediate scores are strongly under-represented, with only 3 patients in the Moderate class within the test set. The four-class ordinal analysis is therefore exploratory and statistically unreliable for the middle classes, and its behavior should be confirmed on larger and more balanced cohorts.

In addition, the decision thresholds t_1, t_2, t_3 used for the four-class ordinal classification were selected on the test set itself. A more rigorous assessment would require an additional external test set on which to validate these thresholds. The present four-class results should therefore be regarded as a potential, preliminary evaluation that nonetheless provides useful insights for future work.

Furthermore, only the lumbar sensor was used. Even if this placement is the

most clinically motivated and the most common, sensors on the wrists, shanks, or chest may capture complementary aspects of postural impairment that are not visible in the lower-back signal alone, and could therefore improve performance.

Finally, medication state was not uniformly recorded across the three datasets. Since dopaminergic therapy directly affects motor symptoms in PD, this introduces an uncontrolled source of variability that feature-level domain adaptation cannot fully resolve.

Chapter 6

Conclusion and Future Work

This study addressed the problem of cross-dataset classification of postural instability in PD from a single wearable sensor placed on the lower back, a symptom that is clinically very relevant but highly under-explored in the literature compared to gait or tremor. The work was motivated by three gaps in the existing studies: the absence of cross-dataset evaluation in IMU-based postural stability assessment, the separate use of handcrafted and latent features from deep networks, and the lack of domain adaptation methods for this specific task. To address them, data from three heterogeneous clinical datasets, WearGait-PD, Kiel, and OmniaPark, were merged for a total of 149 PD patients recorded in the United States, Germany, and Italy. Only the triaxial accelerometer and gyroscope signals from one lumbar IMU were leveraged, and the MDS-UPDRS Item 3.12 score was used as the ground-truth label. First, a pre-processing pipeline was applied to align axes, units, and sampling rate between the different datasets and to address class imbalance through an adaptive windowing strategy. Then, two encoder architectures, a supervised CNN-GRU classifier and a self-supervised masked autoencoder, were trained to extract window-level latent representations. These were aggregated at the patient level and concatenated with 14 handcrafted clinical descriptors into a single 46-dimensional feature vector and optionally aligned using CORAL or MMD domain adaptation. Finally, subjects are classified into Stable (MDS-UPDRS Item 3.12 score 0) or Unstable (score ≥ 1) by four classical machine learning models.

The results show that the cross-dataset approach is feasible. Even without any domain adaptation, the best configurations reached a ROC-AUC of around 0.85 from a single lumbar sensor, confirming that the lower-back inertial signal already

carries enough information to separate stable from unstable patients across different sites, hardware, and protocols. Domain adaptation improved this further: CORAL achieved the best performance, leveraging latent features from the supervised encoder and logistic regression, up to a ROC-AUC of 0.97, while MMD produced smaller but more consistent gains across configurations. Both methods were effective in addressing domain shift, and domain accuracy dropped from 0.88–0.97 on the raw features to 0.43–0.48 after correlation alignment and to 0.66–0.73 after mean-shift.

A relevant finding is that the self-supervised autoencoder, trained without any clinical label, matched the supervised encoder almost exactly (best ROC-AUC of 0.95 against 0.97), which indicates that meaningful representations of postural dynamics can be learned from unlabeled inertial data. This is especially relevant given the scarcity of labeled clinical data.

The ablation across feature subsets further clarified the contribution of each component: handcrafted descriptors alone were the weakest, reaching at most a ROC-AUC of 0.74, latent features alone were substantially stronger, and their combination, after recursive feature elimination, gave the best results, confirming that the two representations are partly complementary. A single-task analysis showed that stance-only and walk-only features performed worse than their combination, suggesting that the two tasks capture complementary aspects of postural stability.

Finally, an exploratory analysis showed that the continuous probability produced by the binary classifier preserves the ordering of the MDS-UPDRS scale, correctly separating the two extremes while confusing only adjacent intermediate scores, suggesting that the learned features implicitly encode disease severity.

These conclusions must be interpreted acknowledging the limitations of this work, such as the limited test size, the single fixed data split, label noise in the clinical ratings, the single-sensor setup and the uncontrolled medication state. These points naturally suggest directions for future research. The most immediate is validation on a larger, independent cohort collected at a new clinical site, which would provide more reliable estimates of generalization. Moreover, it could be useful to test whether sensors on the wrists, shanks, or chest add complementary information. The implicit ordinal structure observed in the predictions suggest an additional regression model on the MDS-UPDRS Item 3.12 score, resulting in clinically more informative outputs. The equivalence between the supervised and self-supervised encoders motivates pre-training the encoder on large collections of unlabeled inertial data before fine-tuning on a small labeled PD set, further reducing the dependence on expensive annotation. Extending the pipeline to longitudinal data, finally, would allow postural instability to be tracked over time, which is of considerable clinical value.

In conclusion, this work demonstrated that cross-dataset postural instability

classification in PD is feasible from a single lumbar IMU, and that domain adaptation is an effective strategy to improve generalization across heterogeneous clinical datasets. The similar performance of the self-supervised and supervised encoders is highly promising, as it suggests future methods may not require manually labeled clinical data.

Bibliography

[1] Anthony J. Anderson, David Eguren, Michael A. Gonzalez, Michael Caiola, Naima Khan, et al. WearGait-PD: An open-access wearables dataset for gait in Parkinson's disease and age-matched controls. *Scientific Data*, 13:306, 2026.

[2] Bruno Andò, Mattia Manenti, Federico Contrafatto, Giovanni Mostile, and Mario Zappia. Advanced analysis of optoelectronic system signals to assess postural behavior in Parkinson's disease. In *2025 IEEE Sensors Applications Symposium (SAS)*, 2025.

[3] Carlo Alberto Artusi, Christian Geroin, Clarissa Pandino, Serena Camozzi, Stefano Aldegheri, Leonardo Lopiano, Michele Tinazzi, and Nicola Bombieri. Dynamic video assessment of axial postural abnormalities in Parkinson's disease: A pilot study. *Movement Disorders Clinical Practice*, 12(5):626–637, 2025.

[4] Marc Bächlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M. Hausdorff, Nir Giladi, and Gerhard Tröster. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2010.

[5] Katherine O. Berg, Sharon L. Wood-Dauphinee, J. Ivan Williams, and Brian Maki. Measuring balance in the elderly: Validation of an instrument. *Canadian Journal of Public Health*, 83(Suppl. 2):S7–S11, 1992.

[6] Janusz W. Błaszczyk, Robert Orawiec, Dorota Duda-Kłodowska, and Grzegorz Opala. Assessment of postural instability in patients with Parkinson's disease. *Experimental Brain Research*, 183(1):107–114, 2007.

[7] Nicolaas I. Bohnen, Martijn L. T. M. Müller, Natalia Zarzhevsky, Robert A. Koeppe, Carolyn W. Bogan, Michael R. Kilbourn, Kirk A. Frey, and Roger L. Albin. Leukoaraiosis, nigrostriatal denervation and motor symptoms in Parkinson's disease. *Brain*, 134(8):2358–2365, 2011.

[8] Luigi Borzì et al. FoG-STAR: Freezing of gait severity, tasks, activities, and ratings dataset. *Scientific Data*, 13:305, 2026.

[9] Luigi Borzì, Silvia Fornara, Federica Amato, Gabriella Olmo, Carlo Alberto Artusi, and Leonardo Lopiano. Smartphone-based evaluation of postural stability in Parkinson's disease patients during quiet stance. *Electronics*, 9(6):919, 2020.

[10] Luigi Borzì, Luis Sigcha, Daniel Rodríguez-Martín, and Gabriella Olmo. Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. *Artificial Intelligence in Medicine*, 135:102459, 2023.

[11] Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, Stephen H. Friend, and Andrew D. Trister. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, 3:160011, 2016.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Captiks Srl. Movit system. https://www.captiks.com/en/products/movit-system, 2024.

[14] Dongzhou Cheng, Lei Zhang, Shuoyuan Wang, Hao Wu, Lutong Qin, and Aiguo Song. MaskCAE: Masked convolutional autoencoder via sensor data reconstruction for self-supervised human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(5):2687–2698, 2024.

[15] Kelvin L. Chou, Jorge Zamudio, Peter Schmidt, Catherine C. Price, Sotirios A. Parashos, Bastiaan R. Bloem, Kelly E. Lyons, Chadwick W. Christine, Rajesh Pahwa, Ivan Bodis-Wollner, Wolfgang H. Oertel, Oksana Suchowersky, Michael J. Aminoff, Irene A. Malaty, Joseph H. Friedman, and Michael S. Okun. Hospitalization in Parkinson disease: A survey of National Parkinson Foundation centers. *Parkinsonism & Related Disorders*, 17(6):440–445, 2011.

[16] Moo K. Chung. Introduction to logistic regression, 2020.

[17] Veronica Cimolin et al. Computation of gait parameters in post-stroke and Parkinson's disease: A comparative study using RGB-D sensors and optoelectronic systems. *Sensors*, 22(3):824, 2022.

[18] Gastone Ciuti, Leonardo Ricotti, Arianna Menciassi, and Paolo Dario. MEMS sensor technologies for human centred applications in healthcare, physical activities, safety and environmental sensing: A review on research activities in Italy. *Sensors*, 15(3):6441–6468, 2015.

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[20] Franco Franchignoni, Emilia Martignoni, Giorgio Ferriero, and Claudia Pasetti. Balance and fear of falling in Parkinson's disease. *Parkinsonism & Related Disorders*, 11(7):427–433, 2005.

[21] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[22] Selene Gallo, Mariachiara Patera, Gabriele Imbalzano, Alessandro Zampogna, Marco Ghislieri, Umberto Mosca, Gianluca Amprimo, Luigi Borzì, Serena Cerfoglio, Mario Antonio Gazzanti Pugliese, Giacinto Luigi Cerone, Alberto Botter, Veronica Cimolin, Cipriano Claudia, Leonardo Lopiano, Fernanda Irrera, Gabriella Olmo, Antonio Suppa, and Carlo Alberto Artusi. Objective monitoring of axial symptoms in Parkinson's disease (OMNIA-PARK): The study protocol. In *Movement Disorders*, volume 40, 2025. Abstract, 2025 MDS International Congress.

[23] Marco Ghislieri, Laura Gastaldi, Stefano Pastorelli, Shigeru Tadano, and Valentina Agostini. Wearable inertial sensors to assess standing balance: A systematic review. *Sensors*, 19(19):4075, 2019.

[24] Christopher G. Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T. Stebbins, Carl Counsell, Nir Giladi, Robert G. Holloway, Charity G. Moore, Gregor K. Wenning, Melvin D. Yahr, and Lisa Seidl. Movement disorder society task force report on the Hoehn and Yahr staging scale: Status and recommendations. *Movement Disorders*, 19(9):1020–1028, 2004.

[25] Christopher G. Goetz, Barbara C. Tilley, Steven R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B. Stern, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E. Lang, Andrew Lees, Sue Leurgans, Peter A. LeWitt, David Nyenhuis, C. Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A. Teresi, Jacobus J. van Hilten, and Nancy LaPelle. Movement Disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170, 2008.

[26] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[27] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[28] Quanquan Gu, Peiyu Huang, Min Xuan, Xiaojun Xu, Dan Li, Jianzhong Sun, Hualiang Yu, Chao Wang, Wei Luo, and Minming Zhang. Greater loss of white matter integrity in postural instability and gait difficulty subtype of Parkinson's disease. *Canadian Journal of Neurological Sciences*, 41(6):763–768, 2014.

[29] Clint Hansen, Elke Warmerdam, Robbin Romijnders, Markus A. Hobert, and Walter Maetzler. Full-body mobility data to validate inertial measurement unit algorithms in healthy and neurological cohorts. figshare. Dataset, 2022. https://doi.org/10.6084/m9.figshare.20238006.v2.

[30] Marcie Harris-Hayes, Allison W. Willis, Sandra E. Klein, Sylvia Czuppon, Beth Crowner, and Brad A. Racette. Relative mortality in U.S. Medicare beneficiaries with Parkinson disease and hip and pelvic fractures. *The Journal of Bone and Joint Surgery (American Volume)*, 96(4):e27, 2014.

[31] Margaret M. Hoehn and Melvin D. Yahr. Parkinsonism: Onset, progression, and mortality. *Neurology*, 17(5):427–442, 1967.

[32] Jesse V. Jacobs, Fay B. Horak, Van K. Tran, and John G. Nutt. An alternative clinical postural stability test for patients with Parkinson's disease. *Journal of Neurology*, 253(11):1404–1413, 2006.

[33] Yeon-Wook Kim, Kyung-Lim Joa, Han-Young Jeong, and Sangmin Lee. Wearable IMU-based human activity recognition algorithm for clinical balance assessment using 1D-CNN and GRU ensemble model. *Sensors*, 21(22):7628, 2021.

[34] Kimberly Kontson, Ankur Butala, Brittney Muir, Anthony Anderson, Michael Caiola, David Eguren, Michael Gonzalez, Seneca Motely, Kelly Mills, Laureano Moro Velazquez, Najim Dehak, and Emile Moukheiber. Wearables for gait in Parkinson's disease and age-matched controls, 2024. Dataset. https://doi.org/10.7303/SYN52540892.

[35] Antonina Kouli, Kelli M. Torsney, and Wei-Li Kuan. Parkinson's disease: Etiology, neuropathology, and pathogenesis. In Thomas B. Stoker and Julia C. Greenland, editors, *Parkinson's Disease: Pathogenesis and Clinical Aspects*, chapter 1. Codon Publications, Brisbane (AU), 2018.

[36] Yuanrong Luo, Lichun Qiao, Miaoqian Li, Xinyue Wen, Wenbin Zhang, and Xianwen Li. Global, regional, national epidemiology and trends of Parkinson's disease from 1990 to 2021: Findings from the Global Burden of Disease Study 2021. *Frontiers in Aging Neuroscience*, 16, 2025.

[37] Martina Mancini, Fay B. Horak, Cris Zampieri, Patricia Carlson-Kuhta, John G. Nutt, and Lorenzo Chiari. Trunk accelerometry reveals postural instability in untreated Parkinson's disease. *Parkinsonism & Related Disorders*, 17(7):557–562, 2011.

[38] Shenghuan Miao, Ling Chen, and Rong Hu. STMAE: Spatial-temporal masked autoencoder for multi-device wearable human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):172, 2023.

[39] Bhavana Palakurthi and Sindhu Priya Burugupally. Postural instability in Parkinson's disease: A review. *Brain Sciences*, 9(9):239, 2019.

[40] Luca Palmerini, Laura Rocchi, Sabato Mellone, Franco Valzania, and Lorenzo Chiari. Feature selection for accelerometer-based posture analysis in Parkinson's disease. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):481–490, 2011.

[41] James Parkinson. An essay on the shaking palsy. 1817. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 14(2):223–236, 2002.

[42] Parkinson Italia ETS. Conoscere il Parkinson. https://www.parkinson-italia.it/conoscere-il-parkinson/244, 2026.

[43] Ruth M. Pickering, Yvette A. M. Grimbergen, Una Rigney, Ann Ashburn, Gordon Mazibrada, Brian Wood, Peggy Gray, Graham Kerr, and Bastiaan R. Bloem. A meta-analysis of six prospective studies of falling in Parkinson's disease. *Movement Disorders*, 22(13):1892–1900, 2007.

[44] Hajar Rabie and Moulay A. Akhloufi. A review of machine learning and deep learning for Parkinson's disease detection. *Discover Artificial Intelligence*, 5:24, 2025.

[45] Ali H. Rajput, Rajesh Pahwa, Pamela Pahwa, and Alex Rajput. Prognostic significance of the onset mode in parkinsonism. *Neurology*, 43(4):829–830, 1993.

[46] Daniel Rodríguez-Martín, Albert Samà, Carlos Perez-Lopez, Andreu Català, Joan Cabestany, Juan Manuel Moreno-Arostegui, and Angels Bayés. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PLOS ONE*, 12(2):e0171692, 2017.

[47] Luis Sigcha, Luigi Borzì, Federica Amato, Irene Rechichi, Carlos Ramos-Romero, Andrés Cárdenas, Luis Gascó, and Gabriella Olmo. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. *Expert Systems with Applications*, 229:120541, 2023.

[48] Luis Sigcha, Luigi Borzì, Irene Rechichi, Carlos Ramos-Romero, Andrés Cárdenas, Luis Gascó, and Gabriella Olmo. Automatic detection of freezing of gait in Parkinson's disease using IMU signals and deep learning. *Expert Systems with Applications*, 255:124522, 2024.

[49] Kendal Smith, Diego Torricelli, Miguel González-Sánchez, Javier Ricardo Pérez-Sánchez, Elisa Luque-Buzo, Francisco Grandas, Juan Miguel Marín, and Jorge Andrés Gómez-García. Detecting postural instability in Parkinson's disease from IMU-based objective measures. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33:2044–2053, 2025.

[50] Charalampos Sotirakis, Zi Su, Maksymilian A. Brzezicki, Niall Conway, Lionel Tarassenko, James J. FitzGerald, and Chrystalina A. Antoniades. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. *npj Parkinson's Disease*, 9:142, 2023.

[51] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[52] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, pages 443–450, 2016.

[53] Jenna E. Thorp, Peter Gabriel Adamczyk, Heidi-Lynn Ploeg, and Kristen A. Pickett. Monitoring motor symptoms during activities of daily living in individuals with Parkinson's disease. *Frontiers in Neurology*, 9:1036, 2018.

[54] Michele Tinazzi, Christian Geroin, Roongroj Bhidayasiri, et al. Task force consensus on nosology and cut-off values for axial postural abnormalities in Parkinsonism. *Movement Disorders Clinical Practice*, 9(5):594–603, 2022.

[55] Eduardo Tolosa, Alicia Garrido, Sonja W. Scholz, and Werner Poewe. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurology*, 20(5):385–397, 2021.

[56] Elke Warmerdam, Clint Hansen, Robbin Romijnders, Markus A. Hobert, Julius Welzel, and Walter Maetzler. Full-body mobility data to validate inertial measurement unit algorithms in healthy and neurological cohorts. *Data*, 7(10):136, 2022.

[57] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532. PMLR, 2019.

Appendix A

Additional Material

Model	Explicitly set	Relevant defaults
RF	Num. estimators = 50, Max. depth = 3, Min. samples per leaf = 3, Class weight = balanced	Split criterion = Gini, Max. features = sqrt
GBM	Num. estimators = 50, Max. depth = 3	Learning rate = 0.1, Subsample = 1.0
SVM	Kernel = RBF, Probability estimates = on, Class weight = balanced	Regularisation $C = 1.0$, Kernel coeff. γ = scale
LR	Class weight = balanced, Max. iterations = 2000	Regularisation $C = 1.0$, Penalty = L2, Solver = LBFGS
RFE	Num. estimators = 80, Max. depth = 3; Features selected = 15, Step = 5	—

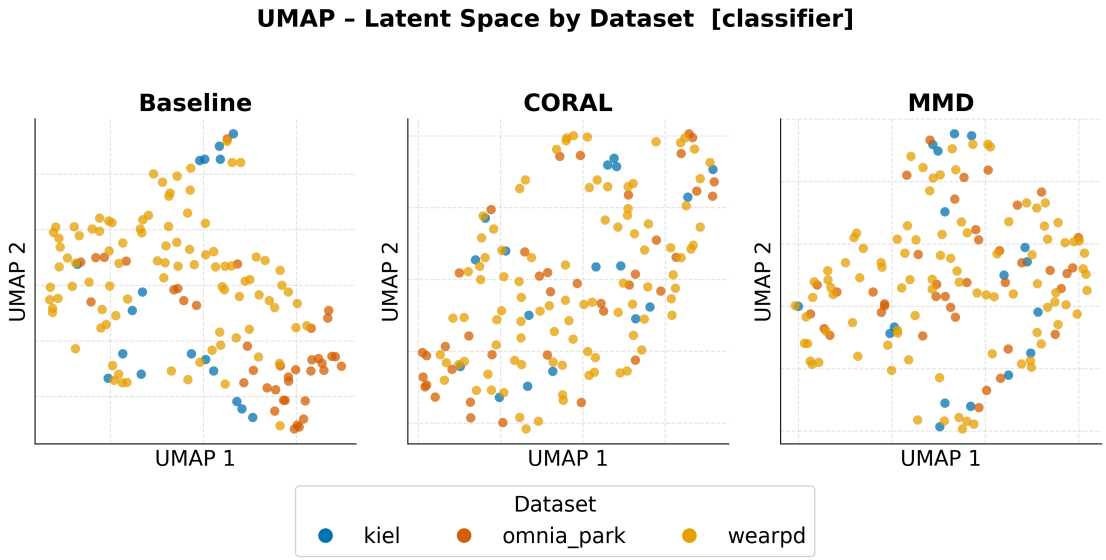
Table A.1: Hyperparameter configuration of the four classifiers and of the RFE feature selector. A fixed random seed (= 42) was used throughout for reproducibility. Abbreviations: RF = Random Forest; GBM = Gradient Boosting Machine; SVM = Support Vector Machine; LR = Logistic Regression; RFE = Recursive Feature Elimination; RBF = Radial Basis Function; L2 = L2 (ridge) regularisation; LBFGS = Limited-memory Broyden–Fletcher–Goldfarb–Shanno; sqrt = square root of the number of features.

DA	Features	Classifier	*N* feat.	ROC-AUC	Bal. Acc.	Macro F1	Precision	Recall
Baseline	Latent	SVM	32	0.849	0.770	0.762	0.684	0.813
	Combined_RFE	LR	15	0.832	0.685	0.681	0.611	0.688
	Combined_RFE	RF	15	0.804	0.676	0.676	0.625	0.625
	Combined	RF	44	0.795	0.685	0.681	0.611	0.688
	Latent	LR	32	0.818	0.662	0.656	0.579	0.688
	Combined	LR	44	0.813	0.676	0.676	0.625	0.625
	Latent	RF	32	0.778	0.707	0.705	0.647	0.688
	Combined_RFE	GBM	15	0.722	0.676	0.676	0.625	0.625
	Latent	GBM	32	0.719	0.676	0.676	0.625	0.625
	Combined	SVM	44	0.810	0.707	0.705	0.647	0.688
	Combined	GBM	44	0.716	0.699	0.700	0.667	0.625
	Handcrafted	RF	12	0.710	0.648	0.632	0.545	0.750
	Combined_RFE	SVM	15	0.815	0.722	0.725	0.714	0.625
	Handcrafted	SVM	12	0.676	0.648	0.632	0.545	0.750
	Handcrafted	LR	12	0.671	0.653	0.652	0.588	0.625
	Handcrafted	GBM	12	0.662	0.628	0.627	0.636	0.438
CORAL	Latent	LR	32	**0.972**	**0.923**	**0.920**	**0.882**	**0.938**
	Combined	LR	44	0.926	0.852	0.861	0.923	0.750
	Combined	RF	44	0.889	0.807	0.809	0.800	0.750
	Latent	RF	32	0.884	0.824	0.815	0.737	0.875
	Combined_RFE	LR	15	0.881	0.824	0.815	0.737	0.875
	Combined_RFE	RF	15	0.855	0.730	0.730	0.688	0.688
	Combined_RFE	SVM	15	0.852	0.750	0.756	1.000	0.500
	Combined_RFE	GBM	15	0.810	0.730	0.730	0.688	0.688
	Latent	GBM	32	0.776	0.676	0.676	0.625	0.625
	Combined	GBM	44	0.787	0.730	0.730	0.688	0.688
	Combined	SVM	44	0.790	0.688	0.680	1.000	0.375
	Handcrafted	SVM	12	0.736	0.699	0.700	0.667	0.625
	Handcrafted	RF	12	0.625	0.605	0.604	0.583	0.438
	Handcrafted	GBM	12	0.722	0.642	0.635	0.750	0.375
	Latent	SVM	32	0.776	0.665	0.657	0.857	0.375
	Handcrafted	LR	12	0.591	0.520	0.512	0.455	0.313
MMD	Combined_RFE	SVM	15	0.960	0.892	0.892	0.875	0.875
	Latent	SVM	32	0.920	0.807	0.809	0.800	0.750
	Combined	LR	44	0.912	0.798	0.805	0.846	0.688
	Latent	RF	32	0.909	0.798	0.805	0.846	0.688
	Latent	GBM	32	0.909	0.776	0.780	0.786	0.688
	Combined	GBM	44	0.903	0.807	0.809	0.800	0.750
	Combined	RF	44	0.898	0.767	0.774	0.833	0.625
	Combined_RFE	RF	15	0.901	0.753	0.755	0.733	0.688
	Combined_RFE	LR	15	0.830	0.847	0.840	0.778	0.875
	Combined_RFE	GBM	15	0.872	0.790	0.799	0.909	0.625
	Latent	LR	32	0.895	0.776	0.780	0.786	0.688
	Handcrafted	SVM	12	0.741	0.679	0.658	0.565	0.813
	Combined	SVM	44	0.909	0.830	0.835	0.857	0.750
	Handcrafted	GBM	12	0.713	0.622	0.622	0.563	0.563
	Handcrafted	RF	12	0.574	0.554	0.550	0.474	0.563
	Handcrafted	LR	12	0.668	0.582	0.582	0.539	0.438

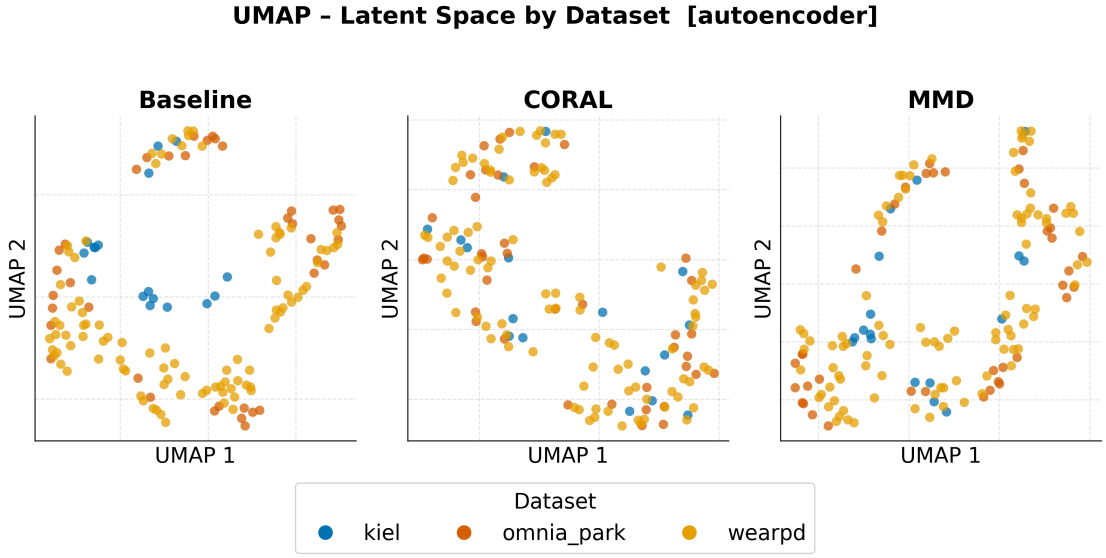
Table A.2: Results of ablation study for the supervised classifier encoder.
Abbreviations: DA = Domain Adaptation; CORAL = Correlation Alignment; MMD = Maximum Mean Discrepancy; *N* feat. = number of features; RFE = Recursive Feature Elimination; RF = Random Forest; GBM = Gradient Boosting Machine; SVM = Support Vector Machine; LR = Logistic Regression; ROC-AUC = Receiver Operating Characteristic Area Under the Curve; Bal. Acc. = Balanced Accuracy; Macro F1 = macro-averaged F1 score.

DA	Features	Classifier	*N* feat.	ROC-AUC	Bal. Acc.	Macro F1	Precision	Recall
Baseline	Combined_RFE	GBM	15	0.869	0.739	0.734	0.667	0.750
	Combined	GBM	44	0.841	0.707	0.705	0.647	0.688
	Latent	SVM	32	0.824	0.778	0.763	0.667	0.875
	Latent	RF	32	0.838	0.716	0.709	0.632	0.750
	Combined_RFE	RF	15	0.838	0.747	0.736	0.650	0.813
	Combined	RF	44	0.830	0.716	0.709	0.632	0.750
	Combined_RFE	SVM	15	0.838	0.716	0.709	0.632	0.750
	Combined_RFE	LR	15	0.815	0.716	0.709	0.632	0.750
	Latent	LR	32	0.801	0.693	0.683	0.600	0.750
	Combined	LR	44	0.801	0.716	0.709	0.632	0.750
	Latent	GBM	32	0.793	0.662	0.656	0.579	0.688
	Combined	SVM	44	0.807	0.716	0.709	0.632	0.750
	Handcrafted	RF	12	0.710	0.648	0.632	0.545	0.750
	Handcrafted	SVM	12	0.676	0.648	0.632	0.545	0.750
	Handcrafted	LR	12	0.671	0.653	0.652	0.588	0.625
	Handcrafted	GBM	12	0.662	0.628	0.627	0.636	0.438
CORAL	Combined_RFE	RF	15	**0.949**	**0.901**	**0.894**	**0.833**	**0.938**
	Latent	LR	32	0.932	0.884	0.890	0.929	0.813
	Latent	RF	32	0.918	0.810	0.790	0.682	0.938
	Combined	RF	44	0.918	0.807	0.809	0.800	0.750
	Combined	LR	44	0.898	0.861	0.864	0.867	0.813
	Combined_RFE	SVM	15	0.898	0.784	0.784	0.750	0.750
	Latent	GBM	32	0.898	0.761	0.759	0.706	0.750
	Combined_RFE	LR	15	0.872	0.770	0.762	0.684	0.813
	Combined	GBM	44	0.830	0.744	0.749	0.769	0.625
	Combined_RFE	GBM	15	0.835	0.699	0.700	0.667	0.625
	Combined	SVM	44	0.832	0.713	0.717	0.750	0.563
	Latent	SVM	32	0.838	0.707	0.705	0.647	0.688
	Handcrafted	GBM	12	0.696	0.619	0.613	0.667	0.375
	Handcrafted	SVM	12	0.659	0.591	0.591	0.533	0.500
	Handcrafted	RF	12	0.614	0.565	0.553	0.556	0.313
	Handcrafted	LR	12	0.585	0.605	0.604	0.583	0.438
MMD	Latent	RF	32	0.906	0.861	0.864	0.867	0.813
	Combined	RF	44	0.901	0.852	0.861	0.923	0.750
	Combined	GBM	44	0.869	0.838	0.838	0.813	0.813
	Combined_RFE	RF	15	0.881	0.838	0.838	0.813	0.813
	Combined_RFE	SVM	15	**0.926**	0.832	0.816	0.714	0.938
	Latent	GBM	32	0.849	0.807	0.809	0.800	0.750
	Combined_RFE	LR	15	0.892	0.798	0.805	0.846	0.688
	Combined_RFE	GBM	15	0.821	0.861	0.864	0.867	0.813
	Latent	SVM	32	0.875	0.753	0.755	0.733	0.688
	Combined	LR	44	0.858	0.668	0.670	0.643	0.563
	Latent	LR	32	0.875	0.730	0.730	0.688	0.688
	Combined	SVM	44	0.869	0.682	0.684	0.727	0.500
	Handcrafted	SVM	12	0.741	0.679	0.658	0.565	0.813
	Handcrafted	GBM	12	0.713	0.622	0.622	0.563	0.563
	Handcrafted	RF	12	0.574	0.554	0.550	0.474	0.563
	Handcrafted	LR	12	0.668	0.582	0.582	0.539	0.438

Table A.3: Results of ablation study for the supervised autoencoder encoder.
Abbreviations: DA = Domain Adaptation; CORAL = Correlation Alignment; MMD = Maximum Mean Discrepancy; *N* feat. = number of features; RFE = Recursive Feature Elimination; RF = Random Forest; GBM = Gradient Boosting Machine; SVM = Support Vector Machine; LR = Logistic Regression; ROC-AUC = Receiver Operating Characteristic Area Under the Curve; Bal. Acc. = Balanced Accuracy; Macro F1 = macro-averaged F1 score.



(a) Supervised classifier encoder.



(b) Autoencoder encoder.

Figure A.1: UMAP projections of the 32-dimensional latent block colored by dataset (WearGait-PD, Kiel, Omnia Park), across the three domain-adaptation variants (Baseline, CORAL, MMD).