Beyond Binary ADHD Classification: Multimodal Phenotyping in a Clinical Psychiatric Cohort

Siyu Li, E. Margaret Perkoff, Lyle Ungar
University of Pennsylvania
Philadelphia, PA, USA
litammy112@gmail.com

Abstract

High-quality automated ADHD assessment has the potential to increase screening rates among individuals in clinical settings, but these methods are far from perfect. Binary classification methods can obscure differences among clinical populations, and computational ADHD studies often evaluate binary diagnostic prediction against neurotypical controls. Such comparisons can obscure which measurements remain useful when ADHD is distinguished from other psychiatric conditions. We address these gaps using the HYPERAKTIV clinical cohort and four measurements: the Conners Continuous Performance Test II (CPT-II), self-report questionnaires, heart-rate variability (HRV), and actigraphy. We examine three aspects of multimodal ADHD assessments: what information different modalities provide for binary ADHD classification, how well multimodal measurements distinguish the clinical groups, and how patterns in the features relate to clinical diagnosis in unsupervised analysis. CPT-II combined with self-report achieved the highest binary AUC (0.908), but was not reliably better than self-report alone under nested cross-validation. In the three-way task, the inattentive group was the hardest to identify. Clustering produced feature profiles that crossed diagnostic groups. These results suggest that adding more modalities does not necessarily improve classification and that feature-space structure does not always align with diagnostic groups.

1 Introduction

Attention-deficit/hyperactivity disorder (ADHD) is often mistaken as a single homogeneous diagnosis, but it is heterogeneous, with presentations that differ in the relative prominence of inattention and hyperactivity-impulsivity (Mostert et al., 2018). In practice, ADHD assessment integrates multiple sources of evidence, including symptom reports, behavioral and cognitive testing, and clinical history, rather than relying on a single feature (Adamou et al., 2024).

There are several types of measurements that capture different aspects of behavior and physiology that can help identify symptoms of ADHD. The Conners' Continuous Performance Test II (CPT-II) (Conners and Sitarenios, 2011) is a standardized computer task that measures sustained attention and response inhibition through omission errors, commission errors, and response-time variability (Huang-Pollock et al., 2012; Fasmer et al., 2016; Egeland et al., 2009). Self-report questionnaires capture difficulties reported by the participant. Heart-rate variability (HRV) measures variation in the time between heartbeats, while actigraphy uses wrist-worn sensors to measure patterns of movement and activity (Ji et al., 2025). Such activity patterns and physiological measures have shown potential for ADHD detection (Shafna and Madhu Kumar, 2024; Ji et al., 2025).

Researchers are leveraging these measurements to automatically screen for ADHD, but the task is often reduced to binary classification: ADHD versus non-ADHD (Hernandez et al., 2023; Pathan et al., 2026). This overlooks differences between ADHD presentations. Additionally, these studies often use neurotypical participants as the comparison group (Casals et al., 2025), making it harder to see which measurements remain useful when ADHD must be distinguished from people with other psychiatric conditions.

We address three questions in this study. First, what information do different measurement modalities provide for binary ADHD classification? Second, how does multiclass classification extend from distinguishing ADHD from psychiatric controls to distinguishing different clinical groups? Third, how do patterns in the measured features relate to diagnostic groups in unsupervised analysis?

We explore these questions using the HYPERAKTIV dataset (Hicks et al., 2021), which contains

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 55–68
November 9, 2026 ©2026 Association for Computational Linguistics

detailed ADHD assessment data from 103 adults. This includes 51 participants with ADHD and 52 psychiatric clinical controls with other mental-health diagnoses. The ADHD participants are further distinguished as inattentive ADHD (ADD) or non-inattentive ADHD groups. This cohort allows us to examine ADHD prediction in a clinically relevant comparison where the non-ADHD group may have other psychiatric conditions, rather than a strictly neurotypical group. [1]

The main contributions of this study are that:

1. **Some modality combinations provide higher observed performance.** CPT-II and self-report achieve the highest observed binary AUC of 0.908 (n = 97), compared with 0.811 for CPT-II and 0.844 for self-report alone.
2. **Multi-class prediction reflects differences between ADHD groups.** In three-way classification, the inattentive group has low recall with every feature set. Even modalities that contribute to strong binary ADHD prediction did not translate well to finer-grained ADHD assessment.
3. **Multimodal measurements reveal structure that extends beyond diagnostic labels.** Across different combinations of CPT-II, self-report, HRV, and actigraphy, clustering identifies groups with distinct measurement patterns. These groups span diagnostic categories rather than simply separating participants according to ADHD status.

2 Related Work

2.1 Multimodal Mental-Health Assessment

Automated methods of mental-health assessment combine information from multiple sources, including cognitive and behavioral measures, self-reports, physiological signals, and wearable sensors. Recent multimodal mental-health systems have combined physiological sensing with language representations (Xu et al., 2026) and integrated transcript text, audio, and facial behavior for clinical assessment (Alavo et al., 2026; Lee et al., 2025). For ADHD, prior work has combined cognitive, behavioral, and physiological measurements, including activity, heart-rate variability (HRV), eye tracking, and EEG (Ji et al., 2025; Castro Ribeiro et al., 2024; Olinic et al., 2025). The Conners' Continuous Performance Test II (CPT-II) is one of the measures used in these computational approaches, as it provides measures of attention, inhibitory control, and reaction-time variability that have been studied in ADHD (Conners and Sitarenios, 2011; Huang-Pollock et al., 2012; Fasmer et al., 2016). Several studies have also compared individual modalities with multimodal combinations (Ji et al., 2025; Wiebe et al., 2024). However, the relative utility of these measurements for distinguishing ADHD from other psychiatric conditions remains an open question. To answer this, we systematically evaluate individual modalities and their combinations in a psychiatric-control cohort.

2.2 ADHD Classification and Heterogeneity

Neuropsychological research has reported heterogeneity in ADHD-related cognitive profiles, including differences in attention, inhibition, delay discounting, and working memory (Mostert et al., 2018). Despite this, computational ADHD assessment is often formulated as a binary classification problem between ADHD and control groups (Hernandez et al., 2023; Loh et al., 2022; Tenev et al., 2014; Gao et al., 2022). Pathan et al. (2026) also applied machine learning to the HYPERAKTIV dataset. While this formulation evaluates overall diagnostic separation, it does not assess whether different ADHD presentations exhibit distinguishable patterns in the measured features. Lin et al. (2023) and other work (Baruah and Kaur, 2024) used machine learning to distinguish inattentive and combined ADHD presentations but lacked psychiatric comparison groups. Other studies used psychiatric comparison groups for ADHD classification (Fasmer et al., 2016) but did not distinguish between presentation types.

Our work builds on this prior body of research by focusing on three-way classification of inattentive, non-inattentive ADHD, and non-ADHD clinical-control groups. We evaluate different features individually and in combination to determine which improve assessment accuracy across these groups.

2.3 Unsupervised Analysis of ADHD Heterogeneity

Supervised classification evaluates predefined diagnostic labels without checking if these labels reflect underlying patterns of variation in the measured features. An unsupervised analysis offers a complementary perspective by identifying patterns without relying on clinical labels. Prior research applied community detection to neuropsycholog-

[1] Code for reproducing the experiments is available here.

ical networks to identify cognitive subgroups in adults with ADHD and healthy controls (Mostert et al., 2018). Such an analysis can also reveal heterogeneity within a diagnostic group, rather than only differences between diagnostic groups.

We apply clustering to combinations of the available modalities and compare the resulting patterns with clinical diagnoses. This analysis explores whether clustering based on multimodal measures organizes participants according to clinical categories or instead reveals patterns that cut across those categories.

3 Dataset and Preprocessing

3.1 Dataset Description

The HYPERAKTIV dataset (Hicks et al., 2021) contains multimodal clinical assessment data from 103 participants aged 17–67, including 51 participants diagnosed with ADHD and 52 psychiatric clinical controls. Participants were assessed in a psychiatric outpatient setting and completed a computerized attention task (CPT-II), while naturalistic activity and physiological data were collected using wearable sensors, and clinical assessments were collected through questionnaires, including ASRS, WURS, MADRS, and HADS, alongside the MINI Plus clinical interview framework. The ADHD group includes 23 individuals with inattentive ADHD (labeled as ADD) and 28 with non-inattentive ADHD. The control group consists of participants with other confirmed psychiatric conditions, including anxiety, bipolar, unipolar depression, substance-use, and related disorders. Thus, it represents a clinically relevant differential-diagnosis setting, instead of a healthy baseline.

3.2 Modalities and Preprocessing

We construct four modality-specific feature blocks based on the HYPERAKTIV data. Table 1 shows the availability of data for each feature type.

CPT-II This block includes twelve raw CPT-II measures, including omissions, commissions, hit reaction time, response-time variability, d', β, perseverations, and block/inter-stimulus-interval effects (Conners and Sitarenios, 2011). Raw values are used for supervised classification to avoid the normative-reference information embedded in pre-normalized T-scores. Standardized CPT-II T-scores are used for clustering so heterogeneous dimensions are on comparable scales.

Property	**Details**	n
Participant overview		
Total cohort	Ages 17–67	103
Non-ADD ADHD	Non-inattentive ADHD	28
ADD	Inattentive presentation	23
Non-ADHD control	Psychiatric comorbidities	52
Per-modality availability		
CPT-II	Raw scores (classification) + T-scores (clustering)	97
Self-report	ASRS, WURS, MADRS, HADS-A, HADS-D	103
HRV (IBI)	Mean NNI, SDNN, RMSSD, pNN50, skewness	80
Actigraphy	Statistical, spectral, and circadian features	85

Table 1: HYPERAKTIV dataset overview and availability of the four modality-specific feature blocks.

Actigraphy Statistical, spectral, and circadian features were extracted from approximately one week of wrist-activity recordings using Actiwatch AW4 devices. These include measures of activity intensity, spectral power, relative amplitude, intra- and inter-daily variability, and timing of daily activity/rest periods (Shafna and Madhu Kumar, 2024).

HRV Five time-domain features were derived from naturalistic ECG-derived inter-beat intervals collected over approximately 20.5 hours using chest-worn Actiheart sensors: mean normal-to-normal interbeat interval (mean NNI), standard deviation of normal-to-normal intervals (SDNN), root mean square of successive differences (RMSSD), percentage of successive intervals differing by more than 50 ms (pNN50), and skewness. Standard artifact rejection and ectopic-beat correction follow established HRV preprocessing procedures (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996).

Self-report Five symptom scales were used in the main experiments: Adult ADHD Self-Report Scale (ASRS; score 0–72) and the Wender Utah Rating Scale (WURS; score 0–100) (Brevik et al., 2020), the Montgomery-Åsberg Depression Rating Scale (MADRS; score 0–60) (Montgomery and Åsberg, 1979), and the Hospital Anxiety and Depression Scale (HADS-A and HADS-D; score 0–21) (Zigmond and Snaith, 1983). Assessments were collected in a psychiatric outpatient setting using the MINI Plus clinical interview framework (Sheehan et al., 1998). Missing numerical values were imputed using median statistics from the training data

within each cross-validation fold, while participants with insufficient questionnaire coverage were excluded. These measures capture ADHD symptoms alongside depression and anxiety burden, which is particularly relevant for the psychiatric-control cohort.

4 Experimental Setup

We conduct three complementary analyses leveraging different combinations of the available feature sets: an ablation study of modalities for binary ADHD-versus-clinical-control classification, a three-way classification of clinical groups, and unsupervised clustering of different modality blocks. For each modality, models use the available data for the corresponding feature set. For models combining multiple modalities, only participants with data available for all included feature sets are used.

4.1 Binary Classification

The first analysis is a binary classification of ADHD versus non-ADHD participants. We compare Logistic Regression (LR), Random Forest (RF), XGBoost, and LightGBM using stratified 10-fold cross-validation. These classifiers were selected to maintain consistency with the original HYPERAKTIV study (Hicks et al., 2021). LR uses L2 regularization and balanced class weights; the tree-based models are included to test whether nonlinear decision boundaries improve over a regularized linear baseline.

We report ROC-AUC as the primary threshold-independent metric and additionally report balanced accuracy, F1, and Matthews Correlation Coefficient (MCC) (Hicks et al., 2022). The observed balanced accuracy is compared to chance odds using a 500-iteration permutation test in which diagnostic labels are randomly permuted while preserving the feature matrix and evaluation procedure.

To test whether the modality ranking above is robust to sample composition, we repeat the binary classification on the common complete-case sample with data available for all modalities ($n = 64$). For each modality block, we use 5×5 repeated stratified cross-validation with a nested inner loop that selects among LR, RF, XGBoost, and LightGBM by validation AUC. We then compute paired AUC differences between each block and the CPT-II + self-report reference block across the repeated folds, and report the 2.5th and 97.5th percentiles of these fold-level differences (Appendix C details the protocol and per-block results).

4.2 Three-Way Classification

We extend the diagnosis classification to three groups: inattentive ADHD, non-inattentive ADHD, and non-ADHD clinical controls. Since the inattentive ADHD group is smaller (23 of 103 participants), we report macro-F1, one-vs.-rest macro-AUC, and balanced accuracy. We use stratified 5-fold cross-validation and evaluate the same modality blocks as in the binary experiment. For each modality block, we report the classifier with the highest cross-validated macro-F1 among LR, RF, XGBoost, and LightGBM; model selection is exploratory, not nested. The three-way task evaluates presentation-sensitive prediction, not clinical diagnosis.

4.3 Unsupervised Cognitive Phenotyping

We use K-Means clustering to group participants based on features from individual modalities and multimodal combinations. PCA is used to visualize the resulting clusters in two dimensions. We then examine whether the resulting groups align with or extend beyond clinically defined categories, as in prior work on cognitive heterogeneity in ADHD (Mostert et al., 2018).

For the clustering analyses, we use a selected subset of the modality combinations to reduce redundancy and preserve interpretable comparisons under limited sample overlap. Certain combinations (e.g., CPT + HRV and CPT + Actigraphy) were omitted from detailed clustering analysis to limit redundancy with the retained combinations. The number of clusters k was selected using silhouette-based model selection (Rousseeuw, 1987). We restricted k to 3–6 because a two-cluster solution can trivially reproduce a coarse split (e.g., ADHD vs. control) and we aimed to look for structure beyond it.

Within-cluster compactness and between-cluster separation were evaluated using the Silhouette score (Rousseeuw, 1987) and Davies-Bouldin (DB) index (Davies and Bouldin, 1979). The alignment between clusters and diagnostic labels was assessed using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and chi-square tests of independence between cluster assignments and diagnostic categories. These analyses are descriptive: agreement with clinical labels does not establish clinical validity, and disagreement may reflect meaningful heterogeneity rather than model failure.

Model	AUC	F1	MCC	Recall	Prec.
Majority	0.50	0.68	0.00	1.00	0.52
LR	**0.91**	**0.79**	**0.61**	**0.80**	**0.81**
RF	0.84	0.74	0.54	0.74	0.80
XGBoost	0.86	0.78	0.57	**0.80**	0.78
LightGBM	0.87	0.78	0.58	0.78	0.80

Table 2: Model comparison on the best modality block (CPT + Self-report, $n = 97$)

Block	n	AUC	bAcc	F1	MCC	p
CPT only	97	0.811	0.745	0.724	0.526	0.002
Self only	103	0.844	0.758	0.751	0.533	0.002
HRV only	80	0.429	0.440	0.354	-0.133	0.798
Act. only	85	0.594	0.597	0.597	0.185	0.146
CPT + Self	97	**0.908**	**0.795**	**0.789**	**0.607**	0.002
CPT + HRV	75	0.788	0.733	0.707	0.476	0.002
CPT + Act.	79	0.726	0.680	0.684	0.371	0.006
CPT+HRV+Self	75	0.840	0.746	0.726	0.518	0.002
All-modality	64	0.769	0.679	0.661	0.362	0.006

Table 3: Binary ADHD vs. non-ADHD classification using Logistic Regression. bAcc = balanced accuracy; Self = self-report; Act. = actigraphy. Bold = best feature block per metric; p = permutation-test p-value.

For supervised classification, all preprocessing steps that estimate parameters from data are performed within the corresponding training folds. Clustering is performed on the complete data available for each modality space. We use fixed feature definitions and scaling procedures for consistency.

5 Experimental Results

5.1 Binary Classification

We first compared Logistic Regression (LR), Random Forest (RF), XGBoost, and LightGBM across all modality blocks (full results in Appendix D); LR achieved the highest AUC on 7 of 9 blocks and is competitive with the best tree-based model on the remaining two. Table 2 shows this comparison on CPT-II + self-report, the strongest modality block, where LR achieves the highest AUC (0.908, $n = 97$), F1, and MCC among the four classifiers. We therefore use LR as the reference classifier for the modality comparison below.

The modality ablation in Table 3 compares LR across modality blocks using each block's complete-case cohort; CPT-II and self-report provide the strongest observed performance. CPT-II+self-report achieves the highest observed AUC (0.908, $n = 97$, permutation $p = 0.002$), exceeding CPT-II alone (AUC=0.811) by 0.097 in this single, non-nested cross-validation. On the common sample (n = 64), the nested estimate AUCs are 0.807 for CPT-II + self-report and 0.784 for CPT-II alone (Table 4), and the paired difference includes zero (Table 9). Self-report alone also performs strongly (AUC=0.844), while adding HRV to CPT-II + self-report achieves AUC=0.840. The physiological-only models perform substantially worse: HRV alone is at or below chance (AUC=0.429, $p = 0.798$), while actigraphy alone (AUC=0.594, $p = 0.146$) is not distinguishable from chance. The all-modality model achieves AUC=0.769 ($n = 64$, $p = 0.006$). Adding HRV or actigraphy to CPT-II does not improve over CPT-II alone.

Block	n	AUC	Fold Range
CPT only	64	0.784	[0.557, 1.000]
Self only	64	0.799	[0.490, 0.952]
HRV only	64	0.385	[0.083, 0.693]
Act. only	64	0.494	[0.248, 0.762]
CPT + Self	64	**0.807**	[0.586, 1.000]
CPT + HRV	64	0.724	[0.430, 0.971]
CPT + Act.	64	0.720	[0.472, 0.895]
CPT+HRV+Self	64	0.786	[0.533, 0.986]
All-modality	64	0.754	[0.540, 0.967]

Table 4: Binary AUC on the common complete-case sample (5×5 repeated nested CV; full metrics in Appendix C). Self = self-report; Act. = actigraphy.

On the common sample (Tables 4 and 9), self-report alone differed from CPT-II + self-report by -0.009 AUC (fold range $[-0.262, 0.152]$). The interval includes zero, so the two cannot be distinguished. HRV-only (difference -0.422, fold range $[-0.774, -0.036]$) and activity-only (difference -0.313, fold range $[-0.667, -0.062]$) remain reliably worse, as their fold ranges lie entirely below zero. All differences are computed as the compared block minus the CPT-II + self-report reference (full robustness protocol and per-block results in Appendix C).

5.2 Three-Way Classification

Table 5 summarizes the performance of the 3-way classification models. The all-modality model had the highest macro-F1 (0.630), followed by self-report questionnaires (macro-F1=0.628). Random Forest, XGBoost, and LightGBM were also evaluated but did not consistently outperform Logistic Regression in this small-cohort setting.

Across all modality combinations, inattentive ADHD (ADD) is consistently the most difficult

Block	Model	Macro F1	Macro AUC	ADD Recall
CPT-II	LGBM	0.517	0.668	0.304
Self-report	LR	0.628	0.777	0.478
HRV only	XGB	0.294	0.458	0.143
Act. only	LR	0.438	0.599	0.435
CPT + Self	LR	0.580	**0.837**	0.348
CPT + HRV	XGB	0.486	0.655	0.238
CPT + Act.	LGBM	0.513	0.717	0.478
CPT + HRV + Self	RF	0.546	0.695	0.333
All-modality	LR	**0.630**	0.758	**0.524**

Table 5: Three-way classification results using the best classifier per modality block with stratified 5-fold cross-validation. Macro-F1/AUC are averaged across classes; ADD Recall denotes recall for the inattentive group.

Block	k	Sil.↑	DB↓	ARI Diag.
CPT-II	4	0.236	1.367	0.050
Self-report	4	0.237	1.262	**0.186**
HRV	3	**0.309**	**1.111**	-0.008
Activity	5	0.135	1.718	-0.005
CPT + Self-report	3	0.161	1.762	0.100
CPT + HRV + Self	4	0.129	1.715	0.095
All-modality	3	0.103	2.175	-0.013

Table 6: Clustering results across modality spaces. Sil. = Silhouette score; DB = Davies–Bouldin index; Diag. = binary ADHD diagnosis. Higher Sil. and ARI indicate stronger cohesion and diagnostic alignment; lower DB indicates better separation. K-Means is used with k selected by silhouette.

class to identify. The all-modality model achieved non-inattentive ADHD recall of 0.667 and non-ADHD recall of 0.710, while ADD recall was 0.524. The lower ADD recall across modality blocks indicates that the inattentive group is difficult to identify from the available feature spaces.

5.3 Unsupervised Cognitive Patterns

Figure 1 and Table 6 summarize clustering results across modality spaces. Internal cluster quality and alignment with diagnostic labels showed different patterns across modalities. Self-report shows the strongest alignment with the binary ADHD diagnosis (ARI=0.186), while HRV produced the highest Silhouette score (0.309) but no diagnostic alignment (ARI=-0.008). These results illustrate that coherent structure within a feature space does not necessarily correspond to clinical diagnostic labels.

CPT-II clustering reveals heterogeneous cognitive patterns CPT-II clustering using K-means clustering ($k = 4$) identified moderate internal structure (Silhouette=0.236; DB=1.367) with limited agreement with binary ADHD diagnosis (ARI=0.050). The four clusters exhibited different ADHD prevalence and captured distinct performance profiles, including low-impairment, mixed-error, and omission-dominant patterns. Rather than reproducing a simple ADHD/non-ADHD partition, the clusters capture variation in cognitive measurements within the clinical cohort. These patterns are exploratory and should not be interpreted as clinically validated subtypes.

Physiological modalities show strong internal structure but weak diagnostic alignment HRV produced the highest Silhouette score (0.309) and a relatively low DB index (1.111), indicating relatively compact and separated clusters within the HRV feature space. However, its ARI with binary ADHD diagnosis was approximately zero (ARI=-0.008), and the association between clusters and diagnosis was not significant ($p = 0.724$). Thus, the relatively compact HRV clusters did not correspond to the binary ADHD diagnosis.

Activity showed weaker internal structure (Silhouette=0.135; DB=1.718) and similarly negligible alignment with the binary ADHD diagnosis (ARI=-0.005; $p = 0.611$). These results show that feature-space structure does not necessarily match diagnostic structure.

Adding modalities does not necessarily improve cluster structure Combining CPT-II with self-report produced three clusters with modest diagnostic alignment (ARI=0.100), while adding HRV yielded four clusters with ARI=0.095. Including all modalities produced the weakest internal clustering structure (Silhouette=0.103; DB=2.175) and no meaningful alignment with binary ADHD diagnosis (ARI=-0.013; $p = 0.226$).

6 Discussion

6.1 Modalities Differ in Discriminative Value

The strongest binary classification model used the combined CPT-II and self-report measures (AUC=0.908; Table 3). The combined model's higher AUC in the full-sample analysis is compatible with complementary information, but the nested common-sample analysis cannot distinguish it from self-report alone; we therefore do not claim complementarity. CPT-II measures standardized attention/inhibition task performance, while questionnaires capture everyday symptoms. Together, they integrate task-based and self-reported mea-

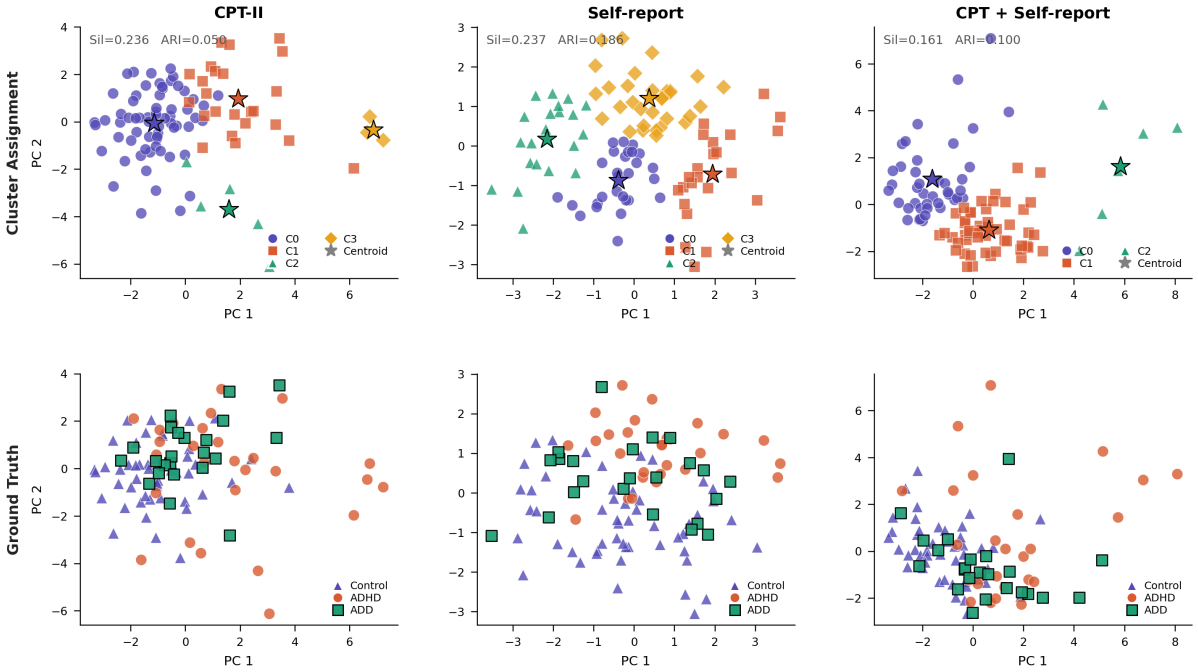


Figure 1: PCA projections of clustering results for CPT-II, self-report, and CPT-II + self-report feature spaces. Top: data-driven cluster assignments; bottom: clinical diagnostic groups. Each point represents one participant.

sures. The strong performance of self-report likely reflects the inclusion of ADHD-specific scales, the ASRS and WURS, which discriminate adults with and without ADHD (Brevik et al., 2020; Du Rietz et al., 2016).

By contrast, HRV and actigraphy contribute less discriminative information when distinguishing ADHD from other psychiatric conditions in this cohort. This does not mean physiological sensing is generally uninformative for ADHD. Rather, physiological and behavioral measures can also reflect other conditions and related factors (Ho et al., 2024). Anxiety and depression can affect autonomic regulation, activity, and sleep rhythms, while psychotropic medications can affect HRV (Alvares et al., 2016). Thus, these signals may be less ADHD-specific in HYPERAKTIV's psychiatric control group than in studies using non-clinical controls (Ji et al., 2025; Shafna and Madhu Kumar, 2024).

6.2 Combining Modalities Does Not Universally Help

Although multimodal models are often expected to improve classification (Ji et al., 2025), adding modalities did not consistently help. CPT+HRV and CPT+actigraphy had lower AUCs than CPT-II alone, while the all-modality model (AUC=0.769) had a lower AUC than CPT-II+self-report (AUC=0.908). This may reflect the addition of less discriminative features and the reduced sample size when requiring complete data across four modalities (97 to 64 participants). The common-sample check (Tables 4 and 9) confirms that HRV-only and actigraphy-only perform reliably worse than the CPT-II+self-report reference in the matched cohort.

6.3 Three-Way Classification Reveals Greater Difficulty for Inattentive ADHD

The three-way experiment exposes a limitation that binary classification can hide. The inattentive ADHD group is consistently harder to classify than non-inattentive ADHD or clinical controls. One possible explanation is that some observed features of inattentive ADHD overlap with those found in other psychiatric conditions. Prior work has shown that adults with inattentive ADHD and adults with mood or anxiety disorders can show overlapping scores on commonly used ADHD measures, while CPT measures may have limited sensitivity and specificity for distinguishing ADHD from other psychiatric disorders (Solanto et al., 2004, 2007).

However, lower recall alone does not explain the hardness. Other differences between the groups, including sample size and clinical characteristics, may also contribute. We therefore interpret the result as evidence that the three-way task is more dif-

ficult than binary ADHD classification, rather than as evidence of a specific boundary between clinical groups. This result is consistent with heterogeneous and overlapping cognitive profiles that are not captured by the available features. Clinicians also integrate longitudinal history, impairment, and developmental trajectory in diagnosis.

6.4 Clustering Reveals Cognitive Patterns

Overall, clustering distinguishes *feature-space structure* from *diagnostic structure*: internally coherent patterns do not necessarily map onto psychiatric categories, and adding modalities may not improve cluster separation or diagnostic alignment.

These results complement the overlapping cognitive profiles in three-way classification. CPT-II clustering reveals distinct performance patterns in omission errors, commission errors, and response variability, but cluster assignments only partially align with diagnoses. Some non-ADHD participants showed more errors or greater variability, while some ADHD participants clustered with controls showing fewer errors and lower variability. This is consistent with prior neuropsychological subgrouping studies reporting overlapping cognitive profiles across clinical and typical populations (Mostert et al., 2018).

Cross-modal clustering also reveals that different modalities produce different divisions of participants. HRV forms relatively compact clusters but shows almost no agreement with ADHD diagnosis, suggesting that structured physiological patterns need not provide strong diagnostic discrimination. Such patterns could be valuable for future work on symptom dimensions, but should not be treated as validated clinical subtypes.

7 Conclusion

We investigated multimodal ADHD assessment in the HYPERAKTIV clinical cohort using CPT-II cognitive-task measures, self-report questionnaires, HRV, and wrist actigraphy. Across binary classification, three-way classification, and unsupervised clustering, the results suggest that using additional modalities may be most useful when they contribute complementary information.

CPT-II combined with self-report has the highest point-estimate AUC in binary classification, though not reliably above self-report alone, while HRV and actigraphy are substantially less discriminative in this cohort. Three-way classification shows that inattentive ADHD is consistently difficult to identify, indicating that finer-grained classification is harder than binary ADHD classification. Finally, unsupervised clustering reveals structured cognitive variation that only partially overlaps with diagnostic categories. These clusters should be viewed as feature-space patterns, not subtypes.

These findings emphasize that multimodal mental-health systems should be evaluated by the clinical context of the comparison group, the incremental value of each modality, and whether patterns in the measured features extend beyond predefined diagnostic categories.

Limitations

The cohort is small (n = 103; 64 with complete data), and approximate cross-validation estimates and intervals mean that all results should be interpreted as exploratory. Larger multimodal datasets covering different ADHD presentations are needed.

Generalizability beyond HYPERAKTIV is also uncertain. The relatively weak HRV and actigraphy results should be interpreted in light of the psychiatric control group, which differs from the participants without known clinical diagnoses used in prior work (Ji et al., 2025; Shafna and Madhu Kumar, 2024). This difference may reflect features of other psychiatric conditions rather than an ADHD-specific signal. Future work should examine modalities across both psychiatric controls and participants without known clinical diagnoses.

Ethical Considerations

This study uses a previously collected and de-identified clinical dataset. Because the cohort comes from a single clinical dataset and lacks systematic demographic, linguistic, or cultural diversity annotations, model performance should not be assumed to generalize across these contexts. The models are intended for exploratory research and should not be used as standalone diagnostic tools. Clustering results represent statistical structure, not clinical presentations. Future applications should include prospective validation, clinician oversight, and evaluation across diverse patient populations.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, and the creators and participants of the HYPERAKTIV dataset for making the data publicly available.

References

Marios Adamou, Muhammad Arif, Philip Asherson, Sally Cubbin, Laurence Leaver, Jane Sedgwick-Müller, Ulrich Müller-Sedgwick, Kobus van Rensburg, and James Kustow. 2024. The adult ADHD assessment quality assurance standard. *Frontiers in Psychiatry*, 15:1380410.

Raymond Alavo, Xinyuan Zhang, Gemza Ademaj, Junhui Cai, Hyeokhyen Kwon, Robert Cotes, Gari D. Clifford, and Ahmed Abbasi. 2026. Dash-M5H: An interactive dashboard for multi-modal, multi-model mental health assessment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 138–147, San Diego, California, United States. Association for Computational Linguistics.

Gail A. Alvares, Daniel S. Quintana, Ian B. Hickie, and Adam J. Guastella. 2016. Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis. *Journal of Psychiatry and Neuroscience*, 41(2):89–104.

Anganabha Baruah and Amanpreet Kaur. 2024. Predicting ADHD subtypes using machine learning on brain imaging data. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–6.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Eirik J. Brevik, Astri J. Lundervold, Jan Haavik, and Maj-Britt Posserud. 2020. Validity and accuracy of the adult attention-deficit/hyperactivity disorder (ADHD) self-report scale (ASRS) and the Wender Utah rating scale (WURS) symptom checklists in discriminating between adults with and without ADHD. *Brain and Behavior*, 10(6):e01605.

Núria Casals, Simon Larsson, and Mikkel Hansen. 2025. Machine learning on a smartphone-based CPT for ADHD prediction. *Frontiers in Psychiatry*, 16:1564351.

Thais Castro Ribeiro, Esther García Pagès, Anna Huguet, Jose A. Alda, Llorenç Badiella, and Jordi Aguiló. 2024. Physiological parameters to support attention deficit hyperactivity disorder diagnosis in children: a multiparametric approach. *Frontiers in Psychiatry*, 15:1430797.

C. Keith Conners and Gill Sitarenios. 2011. Conners' continuous performance test (CPT). In Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 681–683. Springer, New York, NY.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Ebba Du Rietz, Celeste H. M. Cheung, Gráinne McLoughlin, Daniel Brandeis, Tobias Banaschewski, Philip Asherson, and Jonna Kuntsi. 2016. Self-report of ADHD shows limited agreement with objective markers of persistence and remittance. *Journal of Psychiatric Research*, 82:91–99.

Jens Egeland, Stein N. Johansen, and Torill Ueland. 2009. Differentiating between ADHD sub-types on CCPT measures of sustained attention and vigilance. *Scandinavian Journal of Psychology*, 50(4):347–354.

Ole Bernt Fasmer, Kristin Mjeldheim, Wenche Førland, Anita L. Hansen, Vigdis Elin Giæver Syrstad, Ketil J. Oedegaard, and Jan Øystein Berle. 2016. Linear and non-linear analyses of Conners' continuous performance test-II discriminate adult patients with attention deficit hyperactivity disorder from patients with mood and anxiety disorders. *BMC Psychiatry*, 16(1):284.

Runqing Gao, Kesui Deng, and Miaoyun Xie. 2022. Deep learning-assisted ADHD diagnosis. In *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*, pages 142–147, New York, NY, USA. Association for Computing Machinery.

Jonathan Hernandez, Giner Alor-Hernández, Maritza Bustos-López, Laura-Nely Sanchez-Morales, and José Sánchez-Cervantes. 2023. Machine and deep learning algorithms for ADHD detection: A review. In *Artificial Intelligence Applications for Health and Clinical Decision Support*, pages 163–191. Springer.

Steven A. Hicks, Andrea Stautland, Ole Bernt Fasmer, Wenche Førland, Hugo Lewi Hammer, Pål Halvorsen, Kristin Mjeldheim, Ketil Joachim Oedegaard, Berge Osnes, Vigdis Elin Giæver Syrstad, Michael Alexander Riegler, and Petter Jakobsen. 2021. HYPERAKTIV: An activity dataset from patients with attention-deficit/hyperactivity disorder (ADHD). In *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21, pages 314–319, New York, NY, USA. Association for Computing Machinery.

Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979.

Fiona Yan-Yee Ho, Chun-Yin Poon, Vincent Wing-Hei Wong, Ka-Wai Chan, Ka-Wai Law, Wing-Fai Yeung, and Ka-Fai Chung. 2024. Actigraphic monitoring of sleep and circadian rest-activity rhythm in individuals with major depressive disorder or depressive symptoms: A meta-analysis. *Journal of Affective Disorders*, 361:224–244.

Cynthia L. Huang-Pollock, Sarah L. Karalunas, Helen Tam, and Amanda N. Moore. 2012. Evaluating vigilance deficits in ADHD: A meta-analysis of CPT performance. *Journal of Abnormal Psychology*, 121(2):360–371.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Yanqing Ji, Janet Zhang-Lea, and John Tran. 2025. Automated ADHD detection using dual-modal sensory data and machine learning. *Medical Engineering and Physics*, 139:104328.

Chia-Yun Lee, Matúš Pleva, Daniel Hladek, and Ming-Hsiang Su. 2025. Multimodal approaches for stress recognition: A comparative study using the StressID dataset. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 29–34, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

I-Cheng Lin, Shen-Chieh Chang, Yu-Jui Huang, Terry B. J. Kuo, and Hung-Wen Chiu. 2023. Distinguishing different types of attention deficit hyperactivity disorder in children using artificial neural network with clinical intelligent test. *Frontiers in Psychology*, 13:1067771.

Hwee Weng Loh, Chui Pin Ooi, Pritam D. Barua, Erin E. Palmer, Filippo Molinari, and U. Rajendra Acharya. 2022. Automated detection of ADHD: Current trends and future perspective. *Computers in Biology and Medicine*, 146:105525.

Stuart A Montgomery and Marie Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, 134(4):382–389.

Jeanette C. Mostert, Martine Hoogman, A. Marten H. Onnink, Daan van Rooij, Daniel von Rhein, Kimm J. E. van Hulzen, Janneke Dammers, Cornelis C. Kan, Jan K. Buitelaar, David G. Norris, and Barbara Franke. 2018. Similar subgroups based on cognitive performance parse heterogeneity in adults with ADHD and healthy controls. *Journal of Attention Disorders*, 22(3):281–292.

Mara-Simina Olinic, Roland Stretea, and Cristian Chereches. 2025. Wearables in ADHD: Monitoring and intervention—where are we now? *Diagnostics*, 15(18):2359.

Mehzabin Pathan, Kanak Agrawal, Mansi Ahir, Atharv Bardapurkar, Mandar Deshpande, and Anjali More. 2026. Enhancing ADHD diagnosis using machine learning: A study on the HYPERAKTIV dataset. In *Advances in Intelligent Systems and Computing*, pages 359–370. Springer.

Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

V. Shafna and S. D. Madhu Kumar. 2024. Optimal interval and feature selection in activity data for detecting attention deficit hyperactivity disorder. *Computers in Biology and Medicine*, 179:108909.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, and Guy C Dunbar. 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, 59(Suppl 20):22–33.

Mary V. Solanto, Kennett Etefia, and David J. Marks. 2004. The utility of self-report measures and the continuous performance test in the diagnosis of ADHD in adults. *CNS Spectrums*, 9(9):649–659.

Mary V. Solanto, Sharone N. Gilbert, Anu Raj, John Zhu, Sabrina Pope-Boyd, Brenda Stepak, Lucia Vail, and Jeffrey H. Newcorn. 2007. Neurocognitive functioning in AD/HD, predominantly inattentive and combined subtypes. *Journal of Abnormal Child Psychology*, 35(5):729–744.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. 1996. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065.

Aleksandar Tenev, Silvana Markovska-Simoska, Ljupco Kocarev, Jordan Pop-Jordanov, Astrid Müller, and Gerhard Candrian. 2014. Machine learning approach for classification of ADHD adults. *International Journal of Psychophysiology*, 93(1):162–166.

Annika Wiebe, Benjamin Selaskowski, Martha Paskin, Laura Asché, Julian Pakos, Behrem Aslan, Silke Lux, Alexandra Philipsen, and Niclas Braun. 2024. Virtual reality-assisted prediction of adult ADHD based on eye tracking, EEG, actigraphy and behavioral indices: a machine learning analysis of independent training and test samples. *Translational Psychiatry*, 14:508.

Wenxuan Xu, Arvind Pillai, Subigya Nepal, Amanda C. Collins, Daniel M. Mackin, Michael V. Heinz, Tess Z. Griffin, Nicholas C. Jacobson, and Andrew Campbell. 2026. LENS: LLM-enabled narrative synthesis for mental health by aligning multimodal Sensing with language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17458–17481, San Diego, California, United States. Association for Computational Linguistics.

Anthony S Zigmond and Richard P Snaith. 1983. The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67(6):361–370.

A Multimodal Feature Combinations

Table 7 summarizes the feature combinations evaluated in the multimodal experiments. For each combination, participants are included only when data are available for all constituent feature blocks. The number of participants therefore varies across combinations because of incomplete modality recordings.

Feature set	Features	n
CPT-II	12	97
Self-report	5	103
HRV	5	80
Actigraphy	21	85
CPT-II + Self-report	17	97
CPT-II + HRV	17	75
CPT-II + Actigraphy	33	79
CPT-II + HRV + Self-report	22	75
All-modality	43	64

Table 7: Feature sets evaluated in the classification and clustering analyses. For multimodal combinations, n denotes participants with available data for all included modalities.

B Self-Report Instrument Details

Five symptom scales were included in the self-report feature block; MADRS is clinician-administered. Composite scores per instrument were used rather than item-level features to reduce dimensionality and overfitting risk given $n = 103$.

- **ASRS** (Adult ADHD Self-Report Scale, 18 items; score 0–72): a validated screener for adult ADHD symptoms (Brevik et al., 2020).
- **WURS** (Wender Utah Rating Scale, 25 items; score 0–100): retrospective self-report of childhood ADHD symptoms (Brevik et al., 2020).
- **MADRS** (Montgomery-Åsberg Depression Rating Scale, 10 items; score 0–60): clinician-administered depression severity scale, included to capture mood-related comorbidity in the control group.
- **HADS-A** (Hospital Anxiety and Depression Scale — Anxiety, 7 items; score 0–21): measures anxiety symptom severity.
- **HADS-D** (Hospital Anxiety and Depression Scale — Depression, 7 items; score 0–21): measures depression symptom severity.

The ASRS and WURS carry the primary ADHD-specific diagnostic signal, whereas MADRS, HADS-A, and HADS-D capture depressive and anxiety symptom burden that partially distinguishes ADHD from anxiety and mood disorders in the psychiatric control group.

C Common-Sample Robustness Analysis

Table 8 reports the full common complete-case sample results with all metrics. Table 9 reports the paired AUC differences against the CPT-II + self-report reference block, and Table 10 reports the nested-CV model-selection variance underlying these estimates.

Block	n	AUC	(SD)	bAcc	F1	LR %
CPT only	64	0.784	0.125	0.703	0.697	92
Self only	64	0.799	0.128	0.726	0.733	80
HRV only	64	0.385	0.177	0.382	0.409	8
Act. only	64	0.494	0.151	0.529	0.508	32
CPT + Self	64	**0.807**	0.132	0.721	0.715	84
CPT + HRV	64	0.724	0.149	0.665	0.662	68
CPT + Act.	64	0.720	0.127	0.645	0.646	52
CPT+HRV+Self	64	0.786	0.131	0.705	0.702	88
All-modality	64	0.754	0.127	0.667	0.675	56

Table 8: Common complete-case sample results, all metrics (5×5 repeated nested CV; mean over 25 paired folds). bAcc = balanced accuracy; LR % = share of folds where LR was the nested-CV-selected model. Self = self-report; Act. = actigraphy.

Block	ΔAUC	SD	Fold Range
CPT only	−0.023	0.081	[−0.178, 0.124]
Self only	−0.009	0.122	[−0.262, 0.152]
HRV only	−0.422	0.207	[−0.774, −0.036]*
Act. only	−0.313	0.195	[−0.667, −0.062]*
CPT + HRV	−0.083	0.117	[−0.307, 0.100]
CPT + Act.	−0.088	0.142	[−0.372, 0.129]
CPT+HRV+Self	−0.021	0.054	[−0.125, 0.067]
All-modality	−0.053	0.100	[−0.244, 0.102]

Table 9: Paired AUC differences (compared block − CPT + self-report), computed over 25 paired folds on the common complete-case sample ($n = 64$). * Fold range excludes zero. Self = self-report; Act. = actigraphy.

D Full Binary Classification Results

Table 11 reports binary (ADHD vs. non-ADHD) classification performance for all four classifiers (LR, RF, XGBoost, LightGBM) across all nine modality blocks, extending the single-block comparison in Table 2. LR achieves the highest AUC on 7 of 9 blocks and is competitive with the best tree-based model on the remaining two (CPT-II alone: XGBoost 0.83 vs. LR 0.81; CPT-II + Activity: RF 0.78 vs. LR 0.73), supporting the choice

Block	Model	AUC	SD	Count
CPT	LR	0.788	0.126	23
	LGBM	0.738	0.135	2
Self	LR	0.818	0.119	20
	RF	0.720	0.147	5
Act.	XGB	0.546	0.154	12
	LR	0.438	0.164	8
	RF	0.494	0.143	2
	LGBM	0.437	0.084	3
HRV	LGBM	0.478	0.140	14
	RF	0.313	0.098	6
	XGB	0.271	0.224	3
	LR	0.121	0.132	2
CPT + Self	LR	0.832	0.120	21
	XGB	0.639	0.196	2
CPT + Act.	LR	0.726	0.106	13
	RF	0.736	0.123	7
	LGBM	0.651	0.211	4
CPT + HRV	LR	0.773	0.120	17
	RF	0.700	0.116	5
	LGBM	0.485	0.129	3
CPT+HRV+Self	LR	0.809	0.107	22
All-modality	LR	0.767	0.112	14
	XGB	0.716	0.172	6
	RF	0.801	0.111	4

Table 10: Nested-CV model-selection variance by block: AUC and SD for models selected in $\geq$2 of the 25 paired folds; singleton selections (count = 1) omitted for space. Count sums to fewer than 25 where singletons are dropped. Self = self-report; Act. = actigraphy.

of LR as the reference classifier for the modality ablation in Table 3.

E Full Three-Way Classification Results

Table 12 reports three-way (ADD vs. non-inattentive ADHD vs. non-ADHD clinical controls) classification performance for all modality blocks.

The best combination (All, LR) achieved non-inattentive ADHD recall = 0.667 and non-ADHD clinical-control recall = 0.710. The persistent inattentive ADHD recall gap across modality blocks indicates that the inattentive ADHD class is difficult to identify from the available feature spaces.

Table 13 shows per-class performance for the best-performing model. The inattentive ADHD group is the hardest class to separate (F1=0.51), consistent with the recall gap noted above; the two other classes are recovered comparably well (F1$\approx$0.70).

Block	Model	n	AUC	F1
CPT	Maj.	97	0.50	0.68
	LR	97	0.81	0.72
	RF	97	0.81	0.74
	XGB	97	**0.83**	0.71
	LGBM	97	0.81	0.73
Self	Maj.	103	0.50	0.00
	LR	103	**0.84**	0.75
	RF	103	0.82	0.78
	XGB	103	0.79	0.72
	LGBM	103	0.82	0.76
HRV	Maj.	80	0.50	0.00
	LR	80	**0.43**	0.35
	RF	80	0.41	0.39
	XGB	80	0.39	0.35
	LGBM	80	0.39	0.30
Act	Maj.	85	0.50	0.69
	LR	85	**0.59**	0.60
	RF	85	0.41	0.45
	XGB	85	0.39	0.40
	LGBM	85	0.43	0.49
CPT + Self	Maj.	97	0.50	0.68
	LR	97	**0.91**	0.79
	RF	97	0.84	0.74
	XGB	97	0.86	0.78
	LGBM	97	0.87	0.78
CPT + Act	Maj.	79	0.50	0.71
	LR	79	0.73	0.68
	RF	79	**0.78**	0.67
	XGB	79	0.74	0.71
	LGBM	79	0.74	0.70
CPT + HRV	Maj.	75	0.50	0.00
	LR	75	**0.79**	0.71
	RF	75	0.74	0.68
	XGB	75	0.70	0.60
	LGBM	75	0.70	0.65
CPT + HRV + Self	Maj.	75	0.50	0.00
	LR	75	**0.84**	0.73
	RF	75	0.74	0.68
	XGB	75	0.71	0.65
	LGBM	75	0.73	0.70
All-modality	Maj.	64	0.50	0.68
	LR	64	**0.77**	0.66
	RF	64	0.71	0.71
	XGB	64	0.75	0.66
	LGBM	64	0.71	0.66

Table 11: Binary classification results (AUC, F1) for all four classifiers on every modality block. Maj. = majority-class baseline. Bold marks the highest AUC within each block. LR is the highest- or near-highest-AUC classifier on 7 of 9 blocks and is used as the reference classifier for the modality ablation in Table 3.

Block	Mdl	n	mF1	mAUC	recall
CPT	LGBM	97	0.517	0.668	0.304
Self	LR	103	0.628	0.777	0.478
HRV	XGB	80	0.294	0.458	0.143
Act	LR	85	0.438	0.599	0.435
CPT+Self	LR	97	0.580	**0.837**	0.348
CPT+HRV	XGB	75	0.486	0.655	0.238
CPT+ACT	LGBM	79	0.513	0.717	0.478
CPT+HRV+Self	RF	75	0.546	0.695	0.333
All-modality	LR	64	**0.630**	0.758	**0.524**

Table 12: Three-way classification results (5-fold CV). mF1 = macro-F1; mAUC = macro-AUC (OvR); recall = ADD recall. Best per column in bold. All = all-modality; Self = self-report only; Act = activity only; HRV = HRV only.

Class	Prec.	Rec.	F1	n
Non-ADHD clinical control	0.71	0.71	0.71	31
non-inattentive ADHD	0.73	0.67	0.70	12
inattentive ADHD	0.50	0.52	0.51	21
Macro avg	0.65	0.63	0.64	64
Wtd. avg	0.64	0.64	0.64	64

Table 13: Per-class results, best three-way model (all modalities / LR, $n = 64$). Wtd. = weighted. Macro averages are computed on pooled predictions; Table 5 reports fold-averaged macro-F1 (0.630).

F CPT-II Feature-Level Discrimination

Table 14 reports Mann-Whitney U tests with rank-biserial effect sizes (r_{rb}) and Benjamini-Hochberg FDR-corrected p-values (Benjamini and Hochberg, 1995) for all 12 raw CPT-II features (ADHD $n = 50$ vs. non-ADHD $n = 47$). Feature names omit the "Raw Score" prefix. Direction (↑/↓) indicates the higher group.

Eight of twelve features show significant group separation after FDR correction. The observed pattern of higher omission errors, commission errors, and response variability, alongside lower d', is consistent with patterns reported in prior ADHD studies (Huang-Pollock et al., 2012; Fasmer et al., 2016). HitRT does not survive correction.

G CPT-II Clustering: Multi-Scope Analysis

To determine whether clustering reflects inter-group discrimination or intra-group heterogeneity, we applied K-Means across three scopes: full cohort, ADHD-only, and non-ADHD clinical-control-only.

The cluster × ADHD association can only be tested in the full cohort, because the diagnosis label

Feature	r_{rb}	p_{raw}	p_{fdr}	Dir.
Omissions	0.483	<0.001	0.0002	$\uparrow^{**}$
Commissions	0.463	<0.001	0.0003	$\uparrow^{**}$
HitSE	0.420	0.0004	0.0008	$\uparrow^{**}$
VarSE	0.546	<0.001	<0.001	$\uparrow^{**}$
D-Prime	−0.417	0.0004	0.0008	$\downarrow^{**}$
Beta	0.278	0.0184	0.0276	$\uparrow^{*}$
Perseverations	0.509	<0.001	<0.001	$\uparrow^{**}$
HitRTIsi	0.377	0.0014	0.0024	$\uparrow^{**}$
HitRT	−0.147	0.2131	0.2557	↓
HitRTBlock	0.125	0.2903	0.3167	↑
HitSEBlock	0.201	0.0891	0.1189	↑
HitSEIsi	0.074	0.5324	0.5324	↑

Table 14: CPT-II feature discrimination (Mann-Whitney U, $n = 97$). $^{*}p_{\text{fdr}}<0.05$, $^{**}p_{\text{fdr}}<0.01$.

Scope	n	k	Sil.	ARI	χ^2-p
All	97	4	0.236	0.050	0.009**
ADHD only	50	3	0.353	n/a	n/a
Non-ADHD only	47	3	0.162	n/a	n/a

Table 15: CPT-II K-Means at three scopes. ARI against binary ADHD label; χ^2-p tests cluster × diagnosis independence.

is constant within the other two scopes. Within the ADHD-only sample, clustering produced a higher Silhouette score (0.353) than in the non-ADHD sample (0.162). This indicates stronger internal separation in the selected ADHD feature space, but it does not establish that the clusters distinguish inattentive from non-inattentive ADHD. The binary diagnosis label is constant within the ADHD-only sample, so ARI and chi-square tests against that label cannot evaluate presentation-level alignment.

H CPT-II Four-Cluster Cognitive Profiles

Table 16 describes the four clusters recovered by K-Means ($k = 4$, $n = 97$) on CPT-II T-score profiles. T>70 indicates clear impairment (≥2 SD above the normative mean of 50). The profiles range from near-normative performance (C0) through mixed error patterns (C1) to increasingly omission-dominant profiles (C2, C3).

The gradient runs from near-normative profiles (C0) through a mixed impulsive-inattentive pattern (C1) to omission-dominant profiles (C2, C3). C3 is a micro-cluster ($n = 3$), exclusively ADHD, with omission T-scores nearly five SD above the normative mean.

External validation. Kruskal-Wallis tests provide additional evidence of differences across the CPT-II clusters: ASRS differed significantly

Cluster	n	%ADHD	Comm.	Omit.	Profile
C0	65	40%	53.3	50.3	Low impair.
C1	24	71%	65.8	80.5	Mixed
C2	5	80%	45.0	92.6	Omit. dom.
C3	3	100%	64.0	98.7	Severe Omit.

Table 16: CPT-II four-cluster T-score profiles ($k = 4$, $n = 97$). Comm. = commission errors; Omit. = omission errors; Omit. dom. = Omission-dominant; Severe Omit. = Severe omission-dominant. %ADHD = proportion of the cluster with an ADHD diagnosis.

($p = 0.003$; means C0=37.1, C1=48.2, C2=34.2, C3=52.0) and WURS differed significantly ($p = 0.008$; C0=37.5, C1=50.7, C2=31.2, C3=69.0). MADRS did not differ ($p = 0.216$), providing limited evidence that the cluster structure is more closely related to ADHD symptom measures than to the measured depressive symptom burden.

I Per-Modality Clustering Validation

Table 17 summarizes clustering quality and statistical validation across all seven modality configurations.

Modality	k	Sil.	DB	ARI	χ^2-p
CPT	4	0.236	1.367	0.050	0.009
Self	4	0.237	1.262	0.186	<0.001
HRV	3	0.309	1.111	−0.008	0.724
Activity	5	0.135	1.718	−0.005	0.611
CPT+Self	3	0.161	1.762	0.100	0.0002
CPT+HRV+Self	4	0.129	1.715	0.095	0.001
All-modality	3	0.103	2.175	−0.013	0.226

Table 17: Per-modality clustering summary. Sil. = Silhouette (higher better); DB = Davies-Bouldin (lower better); ARI against binary ADHD label; χ^2-p = cluster $\times$ diagnosis test.

Two key observations: (1) HRV achieves the highest silhouette (0.309) but near-zero ARI (-0.008) and non-significant χ^2 ($p = 0.724$) - geometrically compact but diagnostically uninformative clusters. (2) The full modality model achieves the lowest silhouette (0.103) and non-significant alignment ($p = 0.226$), indicating weak cluster structure and diagnostic alignment.

J Cluster Alignment with Binary and Three-Group Diagnostic Labels

Table 18 compares each modality's cluster assignments against binary ADHD diagnosis (ARI-Dx) and the three-group presentation reference (ARI-3G). A positive Δ means the clusters agree slightly more with the three-group labels than with the binary label. Because all ARI values are near zero, we do not interpret Δ as evidence of finer-grained structure.

Modality	k	Sil.	ARI-Dx	ARI-3G	Δ
Self	4	0.237	0.186	0.005	−0.181
HRV	3	0.309	−0.008	0.069	+0.078
Activity	5	0.135	−0.005	0.008	+0.013
CPT+Self	3	0.161	0.100	0.142	+0.043
All-modality	3	0.103	−0.013	0.042	+0.054

Table 18: ARI vs. binary diagnosis (ARI-Dx) and three-group clinical reference (ARI-3G). Δ = ARI-3G − ARI-Dx.

Of the five configurations in Table 18, all except self-report have ARI-3G slightly above ARI-Dx (Δ = +0.013 to +0.078), but all ARI-3G values (0.005–0.142) indicate little agreement with either label set. Self-report shows the opposite pattern (ARI-Dx = 0.186, ARI-3G = 0.005), which is consistent with it separating ADHD from controls overall rather than inattentive from non-inattentive ADHD.

K Comorbidity Composition of the Control Group

Comorbidities	n
0	2
1	16
2	17
3	15
4+	2

Table 19: Comorbidity count distribution for non-ADHD clinical controls ($n = 52$). Conditions include anxiety, bipolar disorder, unipolar depression, and substance use disorder.

Table 19 summarizes comorbidity load among the 52 non-ADHD control participants. Only 2 participants (3.8%) carry zero comorbid diagnoses.

This distribution provides context for interpreting the difficulty of physiological classification: autonomic and circadian dysregulation associated with anxiety (reduced HRV), depression (disrupted circadian rhythms), and bipolar disorder overlaps substantially with the HRV and actigraphy signatures of ADHD. The near-chance performance of HRV-only (AUC=0.429) and actigraphy-only (AUC=0.594) classifiers may reflect shared physiological patterns across conditions rather than a lack of signal in these modalities.