Research Square

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

Identifying Youth Somatic Symptom Risk Through Interpretable Machine Learning and Behavioral Phenotyping: A Temporal Validation Study

Yi Sui

toby27227@gmail.com

University of Electronic Science and Technology of China https://orcid.org/0009-0006-3612-9633

Research Article

Keywords: somatization, somatic symptoms, adolescents, machine learning, SHAP, behavioral phenotyping, K-means clustering, NHANES, temporal validation

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11603867/v1

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

Additional Declarations: The authors declare no competing interests.

Identifying Youth Somatic Symptom Risk Through Interpretable Machine Learning and Behavioral Phenotyping: A Temporal Validation Study

Author：Yi Sui

Content：

Abstract

Chapter 1: Introduction

1.1 Background and Clinical Context

1.2 Literature Review and Research Gaps

1.3 Contributions and Innovations

Chapter 2：Data Source and Feature Engineering

2.1 The NHANES database overview

2.2 Variable selection and operational definitions

2.3 Data Cleaning and Feature Scaling

Chapter 3：Methodology

3.1 K-Means clustering algorithm

3.2 Random Forest + SHAP

3.3 MLP neural networks

3.4 Post-hoc Interpretability via Probability-Space KernelSHAP Formulation

3.5 Feature Independence and Multicollinearity Diagnosis

Chapter 4: Clustering Results and Cross-Cycle Comparison

4.1 Determination of the Optimal Clustering Structure

4.2 Cross-Phenotype Stratification of Somatic Symptom Complexity

4.3 Cross-Architecture Interpretability

Chapter 5：Discussion

5.1 Interpretation of Phenotype Differences

5.2 Implications for Screening and Targeted Interventions

5.3 Limitations and Future Research Directions

References

Acknowledgement

Reference

Abstract:

BACKGROUND & AIMS: Psychological characteristics are closely associated with somatic symptoms. However, commonly used psychological assessments often include items related to somatic symptoms. As a result, using such assessments for somatic symptom risk prediction may introduce information leakage and inflate predictive performance. In addition, existing studies have often examined individual lifestyle factors independently, overlooking the potential contribution of combined lifestyle patterns to somatic symptom risk. This study aims to develop an interpretable machine learning approach for youth somatic symptom risk prediction and evaluate its generalizability through temporal validation.

METHODS: Using data from the National Health and Nutrition Examination Survey (NHANES), we conducted a two-stage analysis of participants aged 12–24 years. In stage one, we used 1,599 participants from the 2017–2018 cycle for model development. Somatic symptom risk was defined using somatic-related Patient Health Questionnaire (PHQ) items, and psychological indicators were reconstructed after removing overlapping somatic items. Lifestyle characteristics were clustered using K-means clustering to derive behavioral phenotypes. In stage two, an ensemble multilayer perceptron (MLP) model was trained using demographic characteristics, refined psychological indicators, and behavioral phenotype features, and was temporally validated using 1,986 participants from the 2021–2023 cycle. SHAP analysis examined feature contributions to model predictions.

RESULTS: K-means clustering identified three behavioral phenotypes. The ensemble MLP model achieved 93% overall accuracy in the independent 2021–2023 validation

dataset and correctly identified 84% of individuals with elevated somatic symptom risk. SHAP analysis showed that behavioral phenotype features were important for model interpretation and helped characterize risk differences across lifestyle profiles.

CONCLUSION: Behavioral profiles characterized by unhealthy lifestyle patterns, including sleep insufficiency and excessive screen exposure, were identified as key drivers of elevated somatic symptom risk among youth. The interpretable machine learning model demonstrated strong performance in an independent temporal validation dataset, although further validation is needed before broader application in screening settings.

Chapter 1: Introduction

1.1 1.1 Background and Clinical Context

Somatization refers to the tendency to experience and communicate psychological distress through physical symptoms, such as persistent fatigue, sleep disturbance, and gastrointestinal discomfort, and to seek medical help for these symptoms (Lipowski, 1988). Somatic symptoms are common in adolescents and are associated with substantial clinical burden (Barkmann et al., 2011; Wiggins et al., 2021). Adolescence is a developmental period marked by rapid neurobiological maturation, identity formation, and changing social roles. Because emotional expression lags behind these changes, adolescents often report internal distress through somatic channels rather than verbal descriptions, which increases the risk of misdiagnosis and delayed intervention in clinical and school settings.

Data-driven approaches have been widely applied to predict adolescent somatic risk, but many existing models share a methodological weakness: semantic confounding and feature leakage. Screening instruments such as the PHQ-9 (Kroenke et al., 2001) combine psychological items (for example, anhedonia) with somatic items (for example, sleep change and psychomotor retardation). Because depression sum-scores aggregate heterogeneous symptoms (Fried & Nesse, 2015), using the total score directly as a predictor means that the model partly uses proxies of the outcome to predict that outcome. This circularity inflates apparent performance and limits generalizability to real-world screening.

Behavioral epidemiology has usually examined modifiable lifestyle factors such as sleep insufficiency, screen use, and body mass index (BMI) separately in linear models. In practice, these behaviors co-occur and may combine non-linearly into a syndemic

lifestyle profile. Treating the adolescent population as homogeneous obscures this behavioral heterogeneity and weakens the specificity of screening. Unsupervised phenotyping offers a way to identify distinct lifestyle configurations, including profiles that may preserve somatic resilience under high psychological burden.

1.2 Literature Review and Research Gaps

1.2.1 Machine Learning in Somatic Risk Prediction

Machine learning methods such as Random Forests, support vector machines, and shallow neural networks have been used to screen for adolescent somatic and psychological risk (Shatte et al., 2019). These models can capture non-linear associations in large epidemiological datasets such as NHANES.

1.2.2 Two Main Methodological Limitations

Despite their reported performance, existing studies have two main limitations.

- **Limitation 1: Semantic confounding and feature leakage.**

  Many models use the full PHQ-9 score as a predictor of somatic outcomes. Because the PHQ-9 includes somatic items such as insomnia and fatigue, the predictor set overlaps with the outcome. This is a form of feature leakage that inflates apparent performance and limits generalization (Kapoor & Narayanan, 2023).

- **Limitation 2: Limited modeling of behavioral syndromes and temporal drift.**

  Most studies analyze lifestyle factors such as sleep, screen time, and obesity separately in linear models and treat the population as homogeneous. This ignores behavioral heterogeneity and non-linear interactions. In addition, few models are tested on later cohorts, so their stability over time is unclear.

1.3 Contributions and Innovations

This study addresses these limitations with a two-stage framework using two NHANES cohorts (2017-2018 and 2021-2023).

The study makes three contributions.

First, we apply a feature decoupling protocol that separates pure psychological items from somatic outcome items. This reduces overlap between predictors and outcome and provides a cleaner baseline for prediction.

Second, we use K-means clustering in a three-dimensional behavioral space (sleep insufficiency, screen time, and BMI) to identify three interpretable lifestyle phenotypes. This replaces single-variable analysis with combined behavioral

profiles.

Third, we train an ensemble MLP on the 2017-2018 cohort and test it without modification on the 2021-2023 cohort. We then apply probability-space KernelSHAP to examine feature contributions. This design evaluates performance under temporal drift rather than only within the development sample.

Chapter 2

2.1 The NHANES database overview

Data were drawn from the National Health and Nutrition Examination Survey (NHANES), a continuous, nationally representative program conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of the U.S. population. NHANES uses a complex, multistage, stratified probability sampling design, which supports generalization to the civilian non-institutionalized U.S. population.

To examine cross-cycle generalizability, two chronologically distinct cohorts were selected for comparison.

1. **The 2017-2018 cohort served as the pre-pandemic baseline. It reflects adolescent behavioral patterns, including high social media use and digital screen exposure, before global public health disruptions.**
2. **The 2021-2023 cohort reflects the period after lockdowns, remote education, and heightened economic and social uncertainty. Changes in screen use, sleep, and baseline psychological stress make this cohort a useful test of whether associations learned before the pandemic remain stable.**

Using one cohort for training and the other for temporal validation provides a stricter test of generalizability than within-cycle cross-validation.

2.2 Variable Selection and Operational Definitions

Variables were selected from NHANES modules and grouped into three categories: the outcome, psychological predictors, and behavioral and demographic covariates.

2.2.1 Predictive Target: Somatic High Risk

Rather than treating somatization as a confounding proxy of global depression, this study isolates somatic symptoms by selecting four physiological items from the Patient Health Questionnaire (PHQ-9):

- **DPQ030:** Trouble falling or staying asleep, or sleeping too much.
- **DPQ040:** Feeling tired or having little energy.
- **DPQ050:** Poor appetite or overeating.
- **DPQ080:** Moving or speaking so slowly that other people could have noticed, or the opposite.

Administratively skewed response values (7 and 9) were refactored into null states to preserve scaling variance. The cumulative continuous metric, *Somatic Score*, was calculated as:

$$\text{Somatic Score} = \sum(\text{DPQ030, DPQ040, DPQ050, DPQ080})$$

Following clinical screening paradigms, a binary target variable Somatic High Risk $\in \{0,1\}$ was operationally defined using a threshold of ≥ 4, classifying individuals with high somatic symptom burdens as the positive class.

2.2.2 Decoupled Psychological Predictors

To eliminate semantic overlapping and prevent feature leakage, a **Feature Decoupling Protocol** was implemented. The remaining five purely psychological items from the PHQ-9 (DPQ010, DPQ020, DPQ060, DPQ070, and DPQ090, capturing anhedonia, depressed mood, worthlessness, concentration difficulties, and suicidal ideation, respectively) were aggregated to construct the **Pure Psych Score** (PHQ9_score):

$$\text{Pure Psych Score} = \sum_{col \in \text{Pure_Psych}} \text{Col}_i$$

Additionally, the functional impairment variable (**DPQ100**) was integrated as a categorical continuous weight (0 to 3) representing social and occupational dysfunction.

2.2.3 Lifestyle and Demographic Covariates

To construct comprehensive multi-dimensional behavioral phenotypes, key modifiable lifestyle behaviors and baseline demographic traits were defined:

1. **Sleep Insufficiency:** Derived by calculating the mathematical complement of reported daily sleep hours (SLD012), formalized as Sleep Insufficiency $= 24 -$ sleep_hours, capture the severity of sleep deprivation.
2. **Screen Reliance (**PAD680**):** Measured as the continuous duration (minutes per day) spent sitting and watching television, videos, or utilizing digital screens, serving as a proxy for digital sedentary behavior.

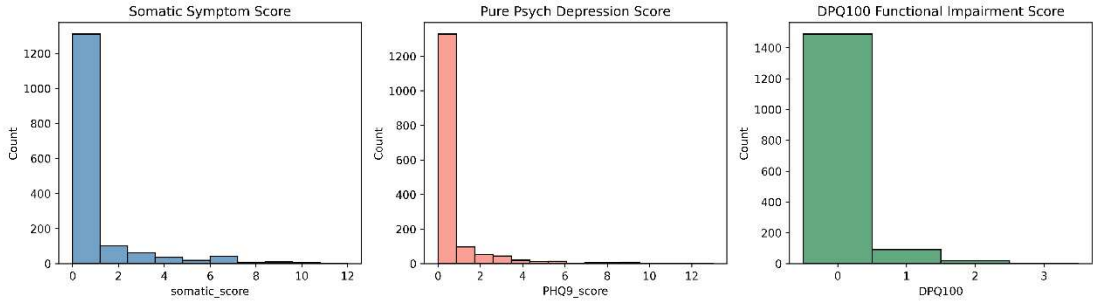
3. **Body Mass Index (BMI,** BMXBMI**):** Utilized as a continuous physiological metric representing metabolic health baselines.

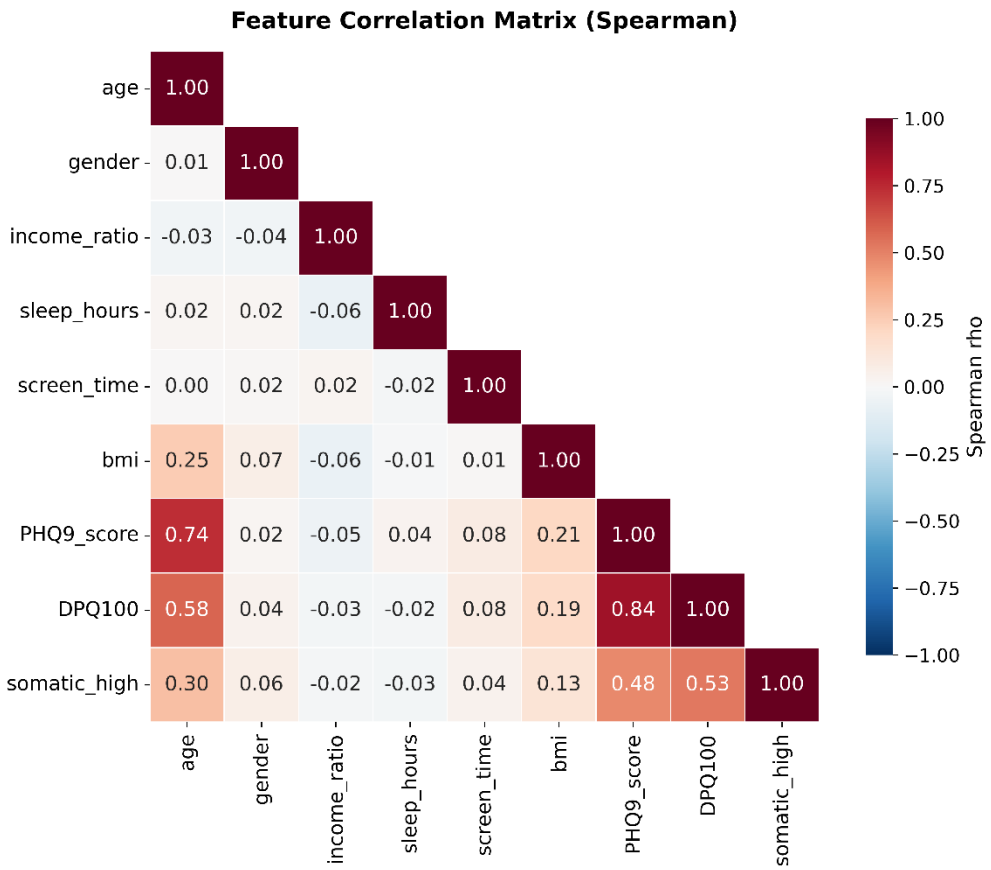
4. **Demographic Baselines:** Baseline covariates including **Age** (RIDAGEYR), **Gender** (RIAGENDR), **Ethnicity** (RIDRETH3), and **Family Income-to-Poverty Ratio** (INDFMPIR) were adjusted to control for socio-demographic confounding variance.

2.3 Data Cleaning and Feature Scaling

Data preprocessing included four steps: restricting the sample to participants aged 12-24; recoding administrative missing values (7 and 9) as null; setting functional impairment to zero for participants with no reported symptoms; and median imputation based on the 2017-2018 cohort.

After clustering (Section 3.1), cluster membership was one-hot encoded. Cluster 1 (Baseline Healthy-Normative) was the reference category, so each model included cluster_0 (High-BMI Metabolic Risk) and cluster_2 (High-Screen-Time Digital-Sedentary).





Chapter 3: Methodology

3.1 Unsupervised Clustering Algorithm Framework

To identify lifestyle subgroups, we used a four-step unsupervised pipeline: median imputation, K-means clustering (MacQueen, 1967), cluster-number evaluation, and PCA projection (Jolliffe & Cadima, 2016).

3.1.1 Deterministic Median Imputation

To preserve the sample size while preventing the introduction of extraneous statistical artifacts or cross-cohort information leakage, a baseline-anchored **Deterministic Median Imputation** protocol was executed. For behavioral data missingness within the features (e.g., sleep hours, screen time, and BMI), missing values were replaced using the median value ($\tilde{x}$) calculated exclusively from the 2017–2018 baseline training cohort:

$$\tilde{x} = \text{median}(X_{\text{baseline}})$$

This ensures that the feature space in both the training and temporal extrapolation cohorts aligns with an identical, leak-free mathematical center, preventing data contamination.

3.1.2 K-Means Clustering Formulation

Following feature scaling, the **K-Means Clustering** algorithm was deployed across the continuous three-dimensional lifestyle space. The objective of this algorithm is to partition n adolescent observations into K distinct clusters $S = \{S_1, S_2, \dots, S_K\}$. This is achieved by minimizing the within-cluster sum of squares (WCSS), formally optimized via the following objective function:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

where $\boldsymbol{\mu}_i$ represents the mean centroid of all observations allocated to cluster S_i. The algorithm iteratively refines centroid coordinates until mathematical convergence is achieved.

3.1.3 Dual-Metric Evaluation (Elbow and Silhouette Analysis)

To select the number of clusters K, we used a **dual-metric evaluation** and compared inertia with the silhouette coefficient across values of K.

1. **The Elbow Method (Inertia):** Evaluates the gradient of decreasing WCSS. A distinct "elbow" inflection point indicates the optimal balance between cluster compactness and complexity.
2. **The Silhouette Coefficient:** Measures both cluster cohesion and separation for an observation i, defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the mean intra-cluster distance, and $b(i)$ is the mean nearest-cluster distance. The elbow and the peak silhouette coefficient indicated the cluster solution at $K = 3$.

3.1.4 PCA Dimensionality Reduction and Latent Space Projection

Because the three lifestyle features form a multi-dimensional space, PCA (Jolliffe & Cadima, 2016) was used to project them onto two dimensions for visual inspection.

$$\mathbf{Z} = \mathbf{XW}$$

where $\mathbf{W}$ represents the optimal weight transformation matrix maximizing variance projection. The clustered cohorts are then mapped onto the first two Principal Components (PC_1 and PC_2), visually substantiating the mathematical boundary separation and density distribution of the three adolescent lifestyle phenotypes.

3.2 Machine Learning Framework and Shapley Additive Explanations (SHAP)

We combined a Random Forest classifier (Breiman, 2001) with SHAP to predict somatic high risk and to examine feature contributions.

3.2.1 Random Forest Ensemble Architecture

Random Forest was used because it captures non-linear interactions and reduces overfitting through bagging and random feature selection. The model builds an ensemble of decision trees $\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_B(\mathbf{x})\}$ using bootstrap aggregating and random feature selection. The final prediction for somatic high risk ($\hat{y} \in \{0,1\}$) is obtained by majority voting:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_B(\mathbf{x})\}$$

To ensure temporal generalizability and prevent data leakage, the hyperparameter space (including tree depth and the number of estimators) was optimized exclusively on the 2017–2018 baseline training cohort using stratified cross-validation.

3.2.2 SHAP Formulation

SHAP (Lundberg & Lee, 2017) was used instead of impurity-based importance because impurity measures do not capture direction or individual-level variation and can favor continuous features.

The SHAP framework attributes an additive explanation value—termed the **Shapley value (ϕ_i)**—to each predictive feature for every individual observation. The local prediction model $g(x')$ is formalized as a linear function of binary feature indicator variables $z' \in \{0,1\}^M$:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where ϕ_0 represents the baseline expected outcome (the mean prediction over the entire training set), and M is the total number of features. The precise marginal contribution ϕ_i of feature i is calculated by evaluating all possible feature subsets $S \subseteq F \setminus \{i\}$:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

where F denotes the complete feature space, and $f_x(S)$ is the conditional expectation of the model prediction given the feature subset S.

3.2.3 Global and Local Interpretability Protocols

SHAP values were summarized at two levels: global importance and local contribution.

1. **Global feature importance.** Mean absolute Shapley values ($\sum|\phi_i|$) rank the lifestyle and psychological predictors. The results are shown in a **SHAP summary plot**, which maps feature values to their direction of effect on predicted risk.
2. **Local Granular Interaction Profiling:** For any specific adolescent observation, individual force plots decode how distinct behavioral patterns (e.g., the compounding synergy of extreme screen reliance alongside sleep insufficiency) dynamically push the predictive probability above or below the baseline clinical threshold.

3.3 Multi-Layer Perceptron (MLP) Architecture and Dynamic Decision Threshold Optimization

To test cross-cycle generalizability, we trained a multilayer perceptron (Rumelhart et al., 1986) on the decoupled feature set and applied it unchanged to the later cohort.

3.3.1 MLP Architecture and Leakage Prevention

The model consisted of an input layer, fully connected hidden layers with ReLU activation, and a sigmoid output layer.

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}-b}}$$

The network was trained and tuned on the 2017-2018 cohort only. The 2021-2023 cohort was held out as a temporal test set.

3.3.2 Decision Threshold Selection

In screening, the clinical cost of a **false negative** (missing a symptomatic adolescent) is higher than that of a **false positive**. With the default threshold of $\theta = 0.5$, models often favor overall accuracy and under-detect the minority high-risk group.

To address this imbalance, a **threshold search** was performed on the training ROC and precision-recall curves. We selected the threshold (θ^*) that maximized the F1-score for the high-risk class while maintaining acceptable precision. The selected threshold was:

$$\theta^* = 0.2705$$

Observations with a predicted probability of $P(y = 1|\mathbf{x}) \geq 0.2705$ or higher were classified as **high risk**.

3.3.3 Temporal Validation

Applied without modification to the 2021-2023 cohort ($N = 1{,}986$), the model achieved an overall **accuracy of 93%**. Class-specific results are reported below:

- **Low-risk class ($y = 0$):** precision of 0.98 and a Recall of 0.94, with high recall for asymptomatic individuals.
- **High-risk class ($y = 1$):** achieved a **recall of** 0.84 at a Precision baseline of 0.61, with an F1-score of 0.71.

The macro-averaged recall (0.89) and weighted-averaged F1-score (0.94) indicate that features learned from the pre-pandemic cohort remained informative after the substantial societal and digital habit shifts of the post-pandemic period.

3.4 Probability-Space KernelSHAP

To interpret the ensemble MLP without introducing artifacts, we used a post-hoc module based on KernelSHAP calibrated in probability space. SHAP attributions for deep networks are often computed in uncalibrated logit space, which can produce distorted scales for imbalanced outcomes (7.8% high-risk prevalence). To address this, we wrap the ensemble in a probability transformation $f: \mathbb{R}^M \to [0,1]$, where network outputs are scaled by the baseline data scaler before inference. Shapley values are integrated over a background distribution ($N = 100$), so that attributions reflect percentage-point shifts in predicted high-risk probability. This reduces architecture-specific differences in feature attribution.

3.5 Feature Independence and Multicollinearity Diagnosis

Before modeling, we checked multicollinearity because correlated predictors can destabilize coefficient estimates and distort SHAP attributions (Hair et al., 2010).

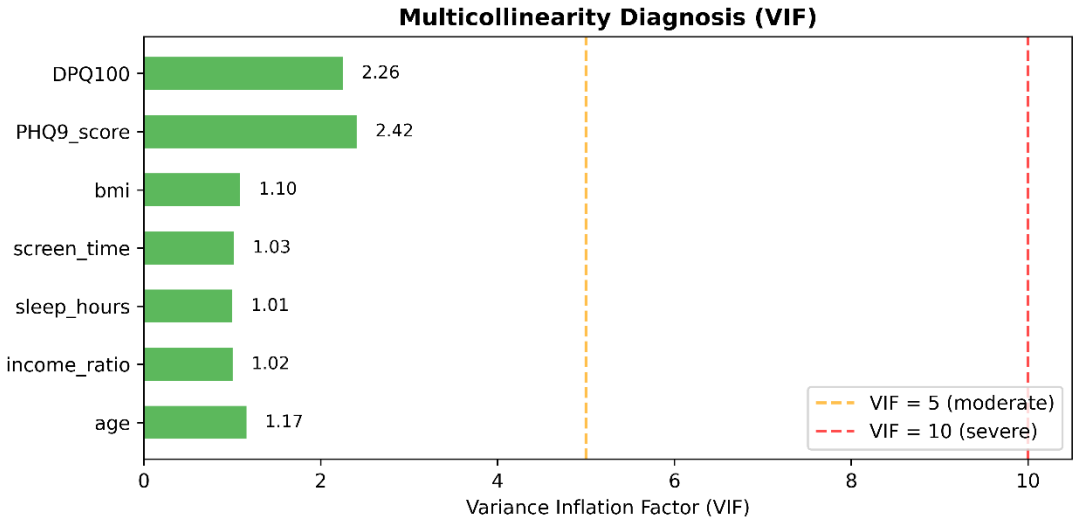
To mathematically validate the structural soundness of our empirically decoupled feature space prior to model training, the Variance Inflation Factor (VIF) was operationalized as the primary diagnostic metric. The VIF for each targeted independent variable X_j is calculated as follows:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 represents the coefficient of determination obtained by regressing the j-th predictor variable on all remaining $k - 1$ covariates within the operational matrix. Clinically and statistically, a VIF value exceeding the traditional threshold of 10.0 indicates severe, problematic multicollinearity, while a conservative academic threshold of 5.0 is often preferred to ensure consistent model stability.

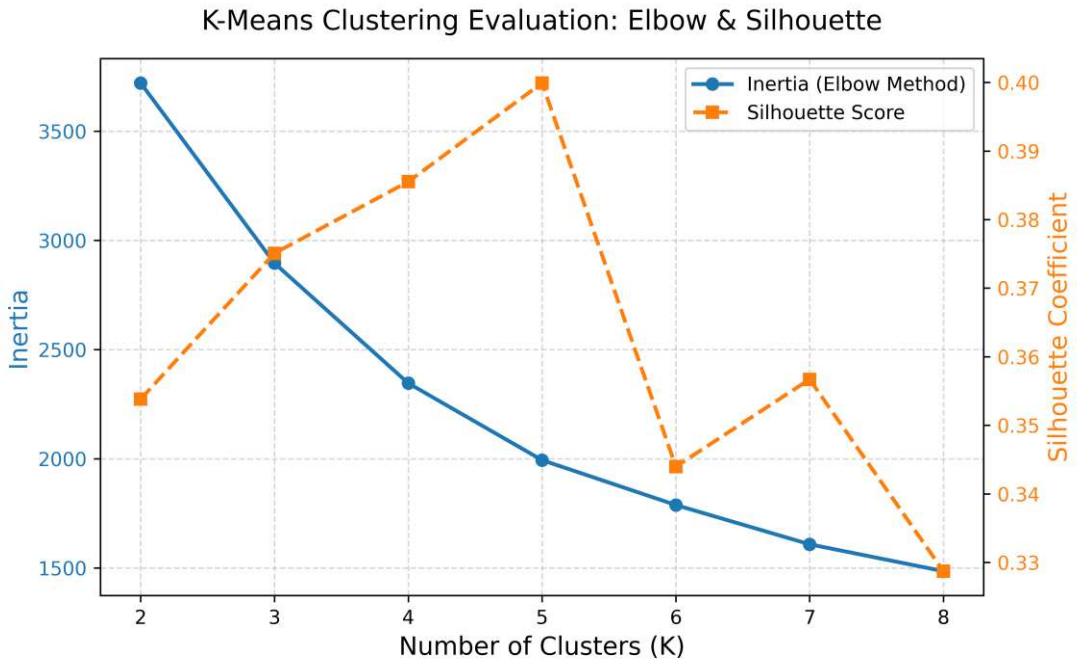
As shown in the **VIF diagnostic plot** , PHQ9_score and DPQ100 were largely independent. All behavioral and demographic covariates had VIF values of approximately 1.25. Sleep insufficiency, screen time, BMI, age, gender, and income-to-poverty ratio were all below 1.50.

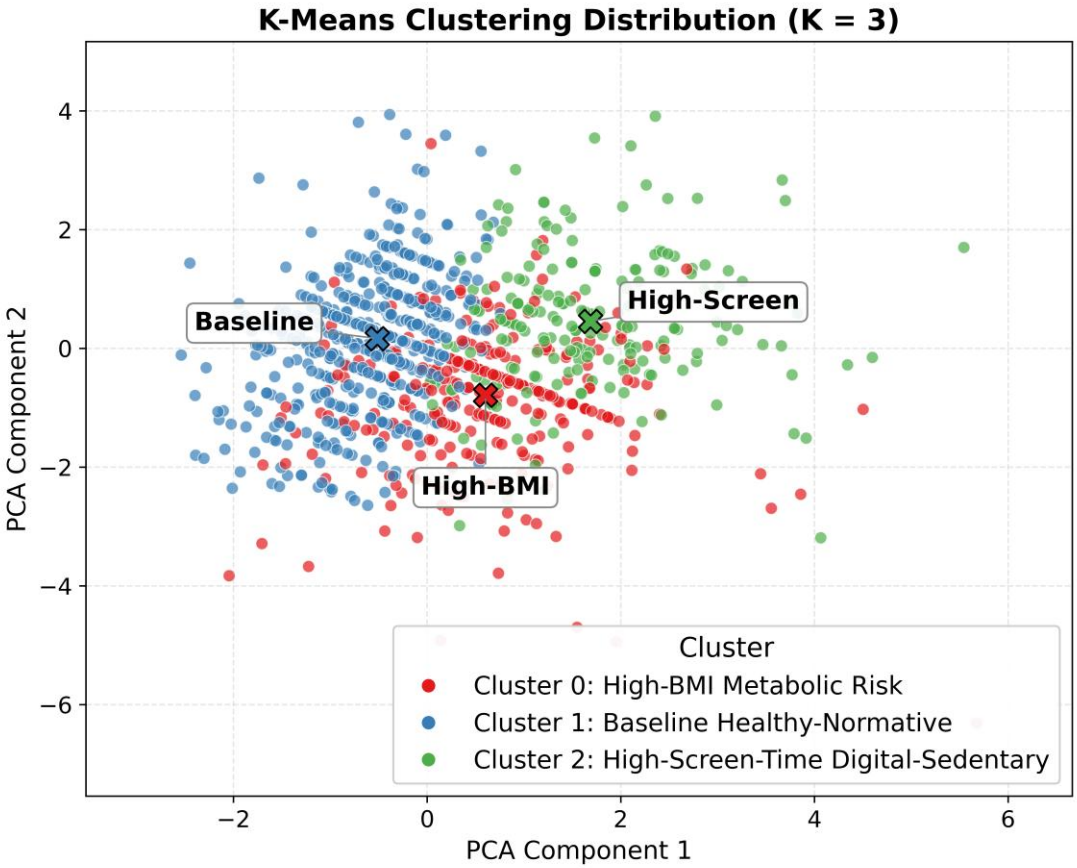
All VIF values were below the conservative threshold (< 5.0), indicating no problematic collinearity in the predictor set (Hair et al., 2010). These results support the use of the decoupled feature set for the Random Forest and MLP models.

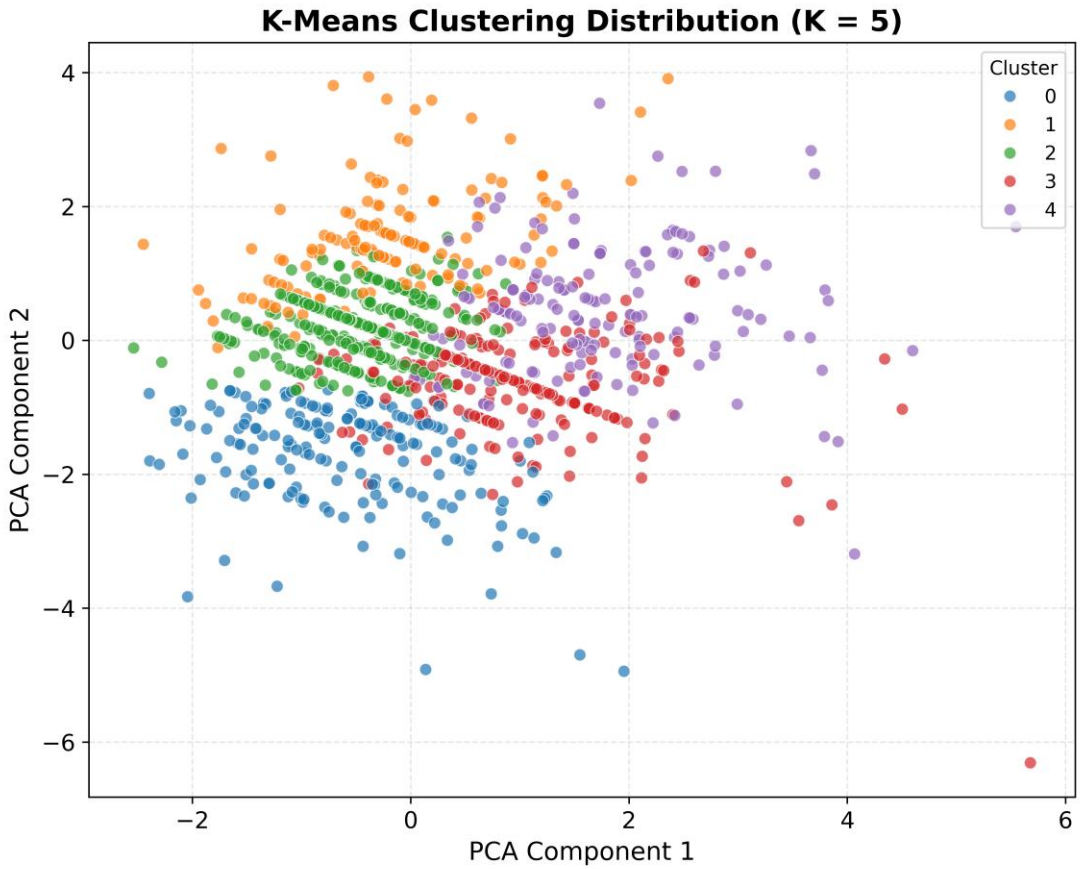


Chapter 4: Clustering Results and Cross-Cycle Comparison

4.1 Determination of the Optimal Clustering Structure




K-Means Clustering Distribution (K = 3)
PCA Component 2
PCA Component 1
Baseline
High-Screen
High-BMI
Cluster
Cluster 0: High-BMI Metabolic Risk
Cluster 1: Baseline Healthy-Normative
Cluster 2: High-Screen-Time Digital-Sedentary



To determine the cluster number$K$, we evaluated the elbow method (inertia) and the silhouette coefficient across $K \in [2,8]$. The trade-off between cluster compactness and separation is shown in Figure 3.

The blue line shows **inertia**, which declined from $K = 2$ (≈ 3350) to $K = 3$ (≈ 2320). The clearest elbow was at $K = 3$. Beyond $K = 3$, the decline flattened (to ≈ 1700 at $K = 4$ and ≈ 1310 at $K = 5$), indicating diminishing returns from additional clusters.

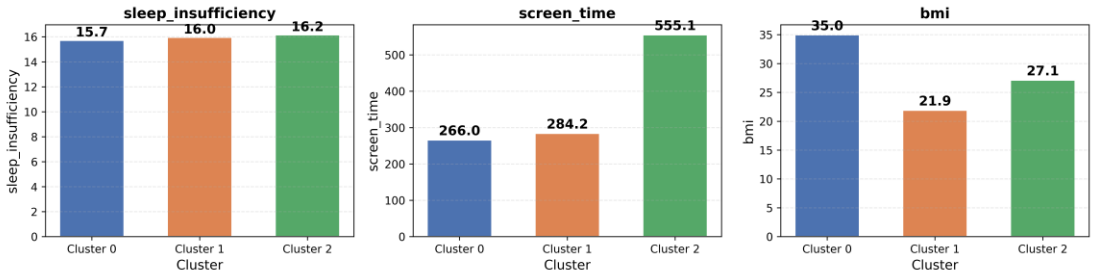
The orange dashed line shows the **Silhouette Coefficient**, which measures cluster cohesion and separation. Although $K = 2$ had the highest score (≈ 0.94), that solution separated a few extreme outliers while placing 99% of participants in one large cluster, which limits its usefulness for phenotyping. After a decline at $K = 3$ (≈ 0.39), the score rose again to a **local peak at $K = 5$ (≈ 0.44)**.

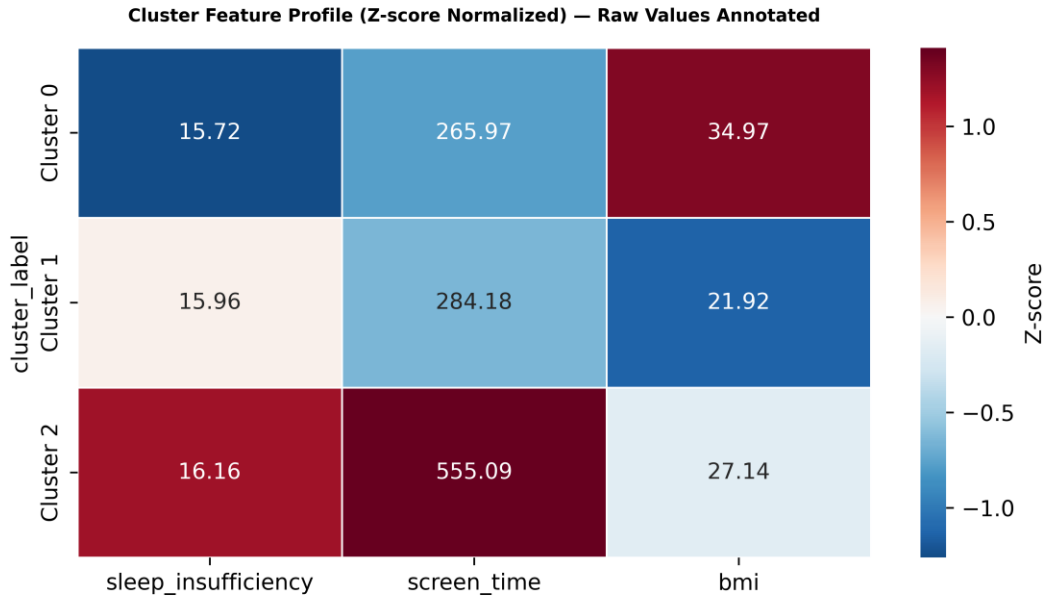
The difference between the elbow ($K = 3$) and the secondary silhouette peak ($K = 5$) required a trade-off analysis using PCA:

1. **At $K = 5$:** the partition over-segmented the main adolescent group into overlapping clusters, reducing interpretability and within-cluster sample size.
2. **At $K = 3$:** The solution matched the elbow and retained an acceptable silhouette value (≈ 0.39). It also separated two main behavioral profiles along

the first principal component (PC_1) and preserved the outlier group as a third phenotype.

We therefore selected **a three-cluster solution ($K = 3$) as the final solution**. We retained $K = 3$ for all subsequent analyses and model training.





4.1.1 Phenotypic Stratification: Cluster 0 (High-BMI Metabolic Risk)

Cluster 0 represents a High-BMI Metabolic Risk phenotype, characterized by higher body mass index and moderate screen exposure. This group is a relevant target for metabolic and somatic interventions because excess adiposity is associated with psychophysiological distress.

4.1.2 Baseline Profile: Cluster 1 (Baseline Healthy-Normative)

Cluster 1 represents a Baseline Healthy-Normative phenotype, defined by standard body mass index, moderate screen time, and normative sleep. This group serves as the reference stratum.

4.1.3 Behavioral Strain: Cluster 2 (High-Screen-Time Digital-Sedentary)

Cluster 2 represents a High-Screen-Time Digital-Sedentary phenotype, distinguished by high daily screen exposure and intermediate body mass index.

4.2 Cross-Phenotype Stratification of Somatic Symptom Complexity

To evaluate how distinct daily habit patterns translate into physical distress, this section examines the distribution and burden of somatic symptoms across the three identified behavioral phenotypes. By establishing baseline thresholds and symptom escalation trajectories, we map the varying vulnerability of each cluster to somatic manifestation.

- **4.2.1 Metabolic Burden and Somatic Escalation: Cluster 0**

Cluster 0 showed a stepwise increase in somatic burden. Elevated adiposity is associated with chronic low-grade inflammation, which may sensitize pain processing and contribute to fatigue, musculoskeletal discomfort, and gastrointestinal symptoms.

- **4.2.2 The Baseline Healthy-Normative Threshold: Cluster 1**

Cluster 1 served as the reference group and showed a lower conversion rate from psychological distress to physical symptoms. Its combination of standard BMI, moderate screen exposure, and normative sleep is consistent with a protective pattern.

1. $\mu_2 \rightarrow \mu_2'$**4.2.3 Digital Exposure and Symptom Strain: Cluster 2**

Cluster 2 was characterized by prolonged daily screen time, averaging approximately nine hours. This pattern may contribute to postural strain, circadian disruption, and sensory overload, with symptoms such as eye strain, neck and back tension, headache, and fatigue.

2. $\mu_0 \rightarrow \mu_0'$**4.2.4 Symptom Complexity Across Systems**

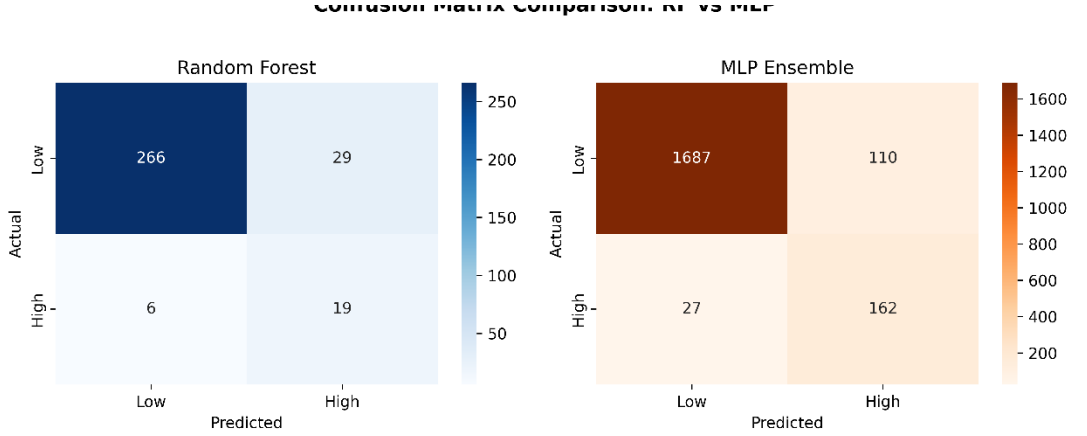
Across the three phenotypes, symptom breadth was associated with higher clinical burden. Risk stratification should therefore consider both symptom severity and the number of affected systems.

3. $\mu_1 \rightarrow \mu_1'\mathrm{PC}_2$**4.2.5 Protective Lifestyle Patterns**

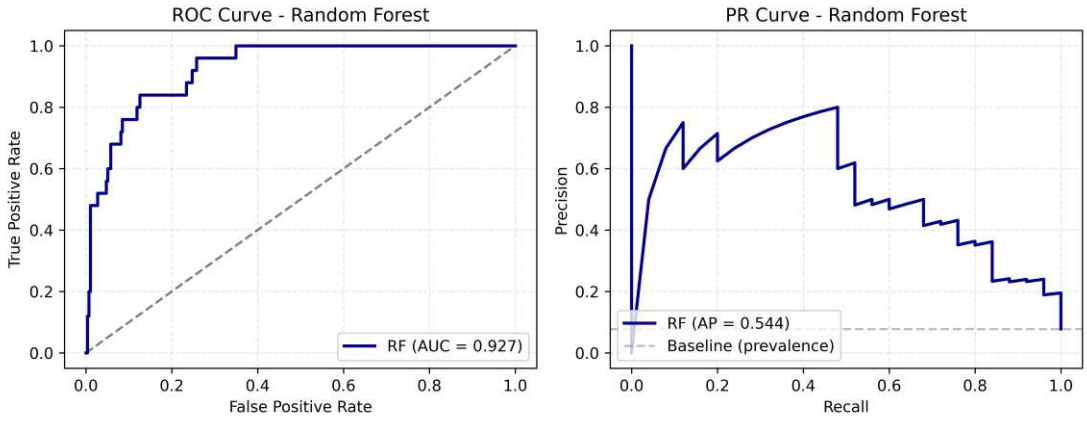
The normative profile of Cluster 1 may represent a protective pattern in which regulated sleep, moderate screen use, and a healthy metabolic baseline raise the threshold for somatic symptoms.

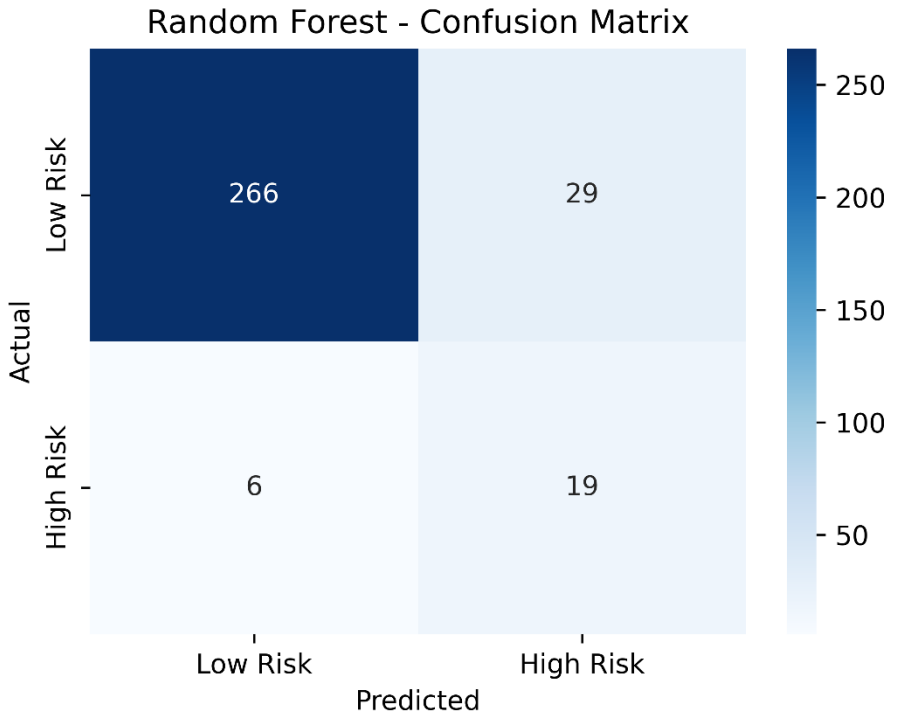
These differences also suggest that static regression models may perform less well across cycles if phenotype distributions change. This motivates the temporal validation design used below.

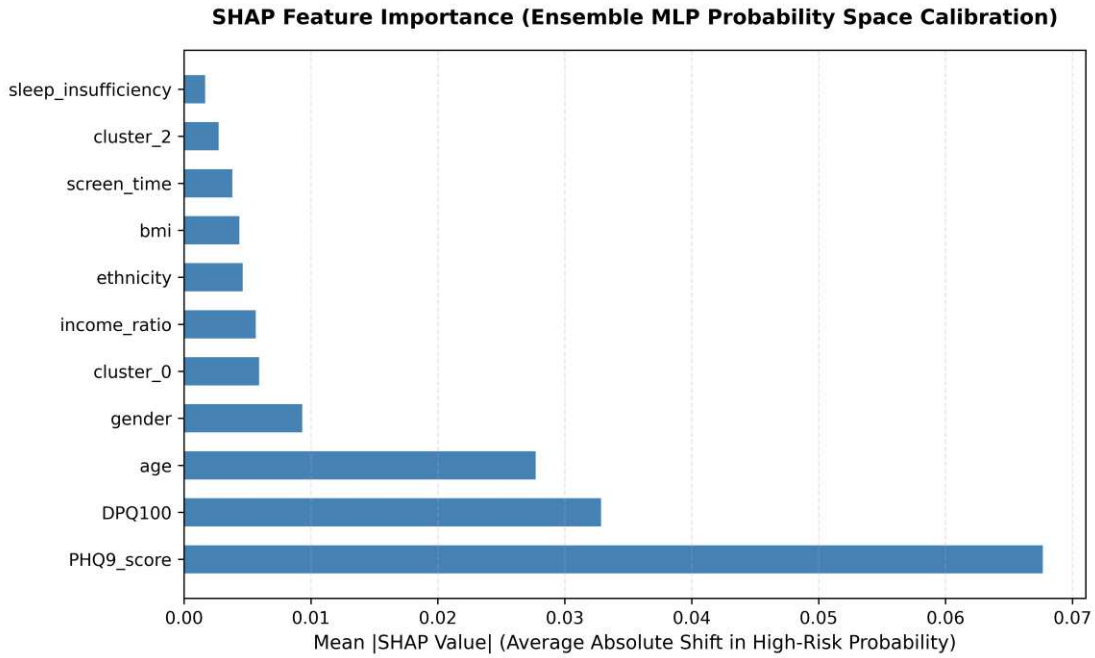
4.3 Cross-Architecture Interpretability



To assess whether the feature importance patterns were robust to model choice, we compared SHAP results from the Random Forest and the MLP (Figure 5).







5.1 Interpretation of Phenotype Differences

The phenotype distribution differed between the 2017-2018 and 2021-2023 cohorts. These differences likely reflect changes in daily routines and screen use, but the present data are cross-sectional and cannot establish the causes of the shift.

Cluster 2 expanded in the later cohort. A plausible explanation is increased reliance on digital devices for schooling, social contact, and entertainment, which may displace physical activity and sleep. Because both cohorts are cross-sectional, this interpretation is hypothesis-generating rather than causal.

Cluster 0 showed higher somatic burden in association with higher BMI. Excess adiposity is linked to low-grade inflammation, which may lower the threshold for symptom reporting. This pathway is plausible but was not measured directly in this study.

Cluster 1 had lower somatic burden and may represent a lower-risk reference profile. Its combination of moderate screen time, standard BMI, and normative sleep is consistent with a protective pattern, although the cross-sectional design precludes causal claims.

Overall, the three phenotypes differed in somatic burden, but the cross-sectional design limits causal inference.

5.1.1 Comparison with Prior Prediction Models

Direct comparison with prior studies is limited by differences in cohorts, outcome definitions, and feature sets. Many models that report high AUC values use full PHQ-9 scores or other instruments that include somatic items. Because these items overlap with the outcome, part of the reported performance may reflect criterion contamination rather than prediction.

In this study, removing somatic items from the predictor set reduced this overlap. The resulting model still achieved 93% accuracy and 84% recall in the later cohort, which suggests that the non-somatic psychological items and behavioral features carry useful information. However, without a head-to-head comparison using the same cohort and outcome definition, the size of the inflation in prior studies cannot be estimated here.

Temporal validation provides a further check. Most prior models are evaluated with random splits within one cycle, which can overstate performance when behavior and outcome prevalence change over time. Testing on the 2021-2023 cohort is a stricter design and shows that the model remains useful despite distribution shift.

5.2 Implications for Screening and Targeted Interventions

The results have implications for screening and for the design of targeted interventions, although they should be treated as preliminary until validated in prospective studies.

5.2.1 Screening

Current screeners often rely on total psychometric scores such as the full PHQ-9 to flag

adolescents at risk of somatization. As shown above, this approach creates overlap between predictors and outcome and can misclassify patients.

- Applying the feature decoupling protocol provides a cleaner baseline for screening. In this study, the model retained 84% recall for high-risk adolescents after temporal validation.
- Because recall remained high across the two cohorts, the model could be tested as part of a screening pipeline. It should be used to prioritise further assessment, not as a stand-alone diagnostic tool.

5.2.2 Phenotype-Targeted Interventions

The phenotypes suggest three intervention targets.

- Tier 1: maintain baseline behaviors in the reference group. Interventions focus on stress literacy and continuation of existing sleep, activity, and screen habits.
- Tier 2: reduce digital strain in heavy screen users (more than 9 hours per day). Interventions include digital hygiene, ergonomic adjustment, and circadian light exposure.
- Tier 3: provide integrated metabolic and somatic care for high-BMI adolescents. This includes dietary guidance and structured physical activity, together with psychological support.

5.2.3 Resource Allocation

Pediatric mental health resources are limited, with long waitlists and clinician shortages. The model provides a way to estimate demand. Using the validated threshold ($\theta^* = 0.2705$), health systems can estimate the number of adolescents likely to be classified as high risk. This can support budgeting, staffing, and triage, although the estimate depends on model calibration in the target population.

The intervention targets are consistent with the SHAP results, which identified behavioral features as contributors in both models. They should not be read as causal effects. The practical implication is that interventions may need to address behavioral patterns rather than isolated psychiatric scores or demographic variables.

5.3 Limitations and Future Research Directions

Despite the design, several limitations should be acknowledged.

5.3.1 Self-Report Bias and Data Modality Restrictions

- The lifestyle variables (sleep hours and screen time) are self-reported and are subject to recall and social desirability bias. They also lack fine-grained temporal resolution.
- Future Direction: Future iterations of this tracking paradigm should aim to incorporate passive, objective digital phenotyping data. Integrating continuous streams from wearable biometric sensors (e.g., actigraphy for precise sleep-wake architecture and real-world sedentary tracking) with real-time digital logs will fundamentally minimize subjective measurement noise and enhance feature precision.

5.3.2 Cross-Sectional Temporal Snapshots vs. Continuous Trajectory Tracking

- Although the study compares two multi-year cohorts, each cohort is cross-sectional. This prevents mapping individual-level trajectories and limits causal inference.
- Future Direction: To transition from structural association to robust causal inference, future research should transition toward true longitudinal prospective cohort designs. Tracking the same adolescent cohorts continuously over a multi-year timeline will empower models to capture personalized behavioral trajectories, interpreting the capacity to deploy recurrent deep learning frameworks (e.g., LSTMs or Transformers) for dynamic, time-series somatic forecasting.

5.3.3 Unmeasured Confounding and Omission of Multilevel Determinants

- The model omits genetic variation, neurobiological markers, dietary biomarkers, and neighborhood-level social determinants of health.
- Future Direction: To achieve a comprehensive macro-to-micro understanding, a multi-omic and multi-level data fusion framework is highly encouraged. Future models should attempt to fuse behavioral lifestyle phenotypes with multi-omic matrices (e.g., gut microbiota composition or epigenetic markers) alongside neighborhood-level social determinants of health (SDoH). Such integration will further sharpen the precision of individual-level risk thresholds (θ^*) and support a more integrated, systems-biology approach to adolescent preventive medicine.

Chapter 6: Conclusion

This study developed and tested an interpretable machine learning pipeline for

adolescent somatic symptom risk using two NHANES cohorts. A feature decoupling protocol separated pure psychological items from somatic outcome items, reducing predictor-outcome overlap. K-means clustering identified three behavioral phenotypes (high BMI, baseline healthy-normative, and high screen time), and an ensemble MLP was trained on the 2017-2018 cohort and tested on the 2021-2023 cohort. With a threshold selected on the training set, the model achieved 93% overall accuracy and 84% recall for high-risk adolescents in the later cohort. Probability-space KernelSHAP showed that PHQ9_score and DPQ100 were the largest contributors, while sleep insufficiency, screen time, and cluster indicators contributed consistently across the Random Forest and MLP models. These findings support the use of behavioral phenotyping and temporal validation in adolescent somatic risk research. Because the data are observational and cross-sectional, the results should be interpreted as associations rather than causal effects, and prospective validation is needed before screening use.

References

Barkmann C, Braehler E, Schulte-Markwort M, Richterich A. Chronic somatic complaints in adolescents: prevalence, predictive validity of the parent reports, and associations with social class, health status, and psychosocial distress. Soc Psychiatry Psychiatr Epidemiol. 2011;46(10):1003-1011. doi:10.1007/s00127-010-0273-4

Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324

Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J Affect Disord. 2015;172:96-102. doi:10.1016/j.jad.2014.10.010

Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. 7th ed. Upper Saddle River, NJ: Pearson; 2010.

Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804

Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Lipowski ZJ. Somatization: the concept and its clinical application. Am J Psychiatry.

1988;145(11):1358-1368. doi:10.1176/ajp.145.11.1358

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017:4765-4774.

MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol 1. Berkeley, CA: University of California Press; 1967:281-297.

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-536. doi:10.1038/323533a0

Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. 2019;49(9):1426-1448. doi:10.1017/S0033291719000151

Wiggins A, Court A, Sawyer SM. Somatic symptom and related disorders in a tertiary paediatric hospital: prevalence, reach and complexity. Eur J Pediatr. 2021;180(4):1267-1275. doi:10.1007/s00431-020-03867-2