frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 08 October 2025
DOI 10.3389/fpsyt.2025.1487894

Check for updates

OPEN ACCESS

EDITED BY
Stephen Aronoff,
Temple University, United States

REVIEWED BY
Caglar Uyulan,
Izmir Kâtip Çelebi University, Türkiye
Long Yin,
Hunan Normal University, China
Youngkyung Moon,
Daejeon University, Republic of Korea

*CORRESPONDENCE
Nina de Lacy
nina.delacy@utah.edu

RECEIVED 28 August 2024
ACCEPTED 08 September 2025
PUBLISHED 08 October 2025

CITATION
de Lacy N, Ramshaw M and Lam WY (2025) Predicting the onset of internalizing disorders in early adolescence using deep learning optimized with AI.
Front. Psychiatry 16:1487894.
doi: 10.3389/fpsyt.2025.1487894

COPYRIGHT
© 2025 de Lacy, Ramshaw and Lam. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Predicting the onset of internalizing disorders in early adolescence using deep learning optimized with AI

Nina de Lacy[1,2]*, Michael Ramshaw[3] and Wai Yin Lam[1,2]

[1]Huntsman Mental Health Institute, Salt Lake City, UT, United States, [2]Department of Psychiatry, University of Utah, Salt Lake City, UT, United States, [3]Center for High Performance Computing, University of Utah, Salt Lake City, UT, United States

Introduction: Internalizing disorders (depression, anxiety, somatic symptom disorder) are among the most common mental health conditions that can substantially reduce daily life function. Early adolescence is an important developmental stage for the increase in prevalence of internalizing disorders and understanding specific factors that predict their onset may be germane to intervention and prevention strategies.

Methods: We analyzed ~6,000 candidate predictors from multiple knowledge domains (cognitive, psychosocial, neural, biological) contributed by children of late elementary school age (9–10 yrs) and their parents in the ABCD cohort to construct individual-level models predicting the later (11–12 yrs) onset of depression, anxiety and somatic symptom disorder using deep learning with artificial neural networks. Deep learning was guided by an evolutionary algorithm that jointly performed optimization across hyperparameters and automated feature selection, allowing more candidate predictors and a wider variety of predictor types to be analyzed than the largest previous comparable machine learning studies.

Results: We found that the future onset of internalizing disorders could be robustly predicted in early adolescence with AUROCs ≥~0.90 and ≥~80% accuracy.

Discussion: Each disorder had a specific set of predictors, though parent problem behavioral traits and sleep disturbances represented cross-cutting themes. Additional computational experiments revealed that psychosocial predictors were more important to predicting early adolescent internalizing disorders than cognitive, neural or biological factors and generated models with better performance. Future work, including replication in additional datasets, will help test the generalizability of our findings and explore their application to other stages in human development and mental health conditions.

KEYWORDS

deep learning, AI, internalizing disorders, adolescence, depression, anxiety, somatic symptom disorder, evolutionary algorithm

Introduction

Depression, anxiety and problematic somatic symptoms (physical symptoms such as headaches and stomachaches) are common mental health issues in adolescence. Often collectively referred to as internalizing disorders, they have been associated with reduced levels of well-being and daily life function, increased risk of self-harm and suicide and are substantial predictors of adult psychopathology (1). Depression and anxiety are among the most common mental illnesses in the population with lifetime prevalence of ~30% and ~20% respectively (2). The incidence of internalizing disorders increases exponentially during the peri-adolescent period, with anxiety having an earlier developmental arc (3). Anxiety disorders emerge during elementary school, with the median age of onset being 11 years of age (yrs) and 75% of lifetime illness occurring by 21 yrs. Major depression cases begin to onset at 11–12 yrs with median onset at 31–32 yrs and 75% of lifetime illness having onset by 44 yrs (4). Problematic somatic symptoms affect up to 40% of youth and increase over peri-adolescence: one third to a half continue to report symptoms as adults with 5-7% in the general population and ~17% in the primary care population meeting criteria as adults for Somatic Symptom Disorder (SSD). (5, 6).

Given the considerable personal, societal and economic burdens associated with internalizing disorders (7–10), there is great interest in identifying specific factors that predict their onset, since evidence suggests that early intervention improves outcomes (11, 12) and reduces resource use (13). Isolating key predictors of internalizing disorders is challenging since they have been associated with a host of different factors from varied domains ranging from biological (neural; genetic; hormonal) and psychological models (fear/threat response) to interpersonal relationship function, parent characteristics, the community environment and wider social determinants of health such as relative poverty. Historically, an important barrier to disambiguating the relative importance of such factors to predicting case onset has been the paucity of appropriate multimodal data in large participant samples. Outside the US, national registries or school system data have been available offering large sample sizes (n>10,000) but these typically lack physiologic information such as neuroimaging data (14–17). An alternative strategy is to combine data from multiple studies offering neuroimaging or genomic data to boost sample size such as the datasets offered by IMAGEN or ENIGMA, though pooling across heterogenous studies may inherently limit features (variables) available for analysis to those that are shared across all studies (18–20). Consequently, to promote comparative discovery at scale, federal and other organizations have recently sponsored the formation of large, longitudinal cohorts collecting a wide variety of multimodal data types with standardized protocols. In peri-adolescence, the flagship initiative of this type is the ongoing population-level ABCD study (n = 11,800) used in the present study (21–23).

Concomitantly, interest has recently grown in applying machine learning (ML) methods to these newly-emerging large-scale population cohorts as ML techniques offer advantages in approaching such high-dimension data. Firstly, they can generate individual-level case predictions from multidimensional data to bridge extant work focused on group-level statistical effects with individual-level discoveries of potential clinical relevance by "providing multivariate signatures that are valid at the single-subject level" (24, 25). Secondly, ML techniques can simultaneously analyze hundreds of candidate predictors and incorporate non-linear relationships among a set of predictors. These properties are relevant to the construction of individual-level models since significant group-level effects may not be useful at the individual level while a feature with low effect size at the group level may prove germane. While a number of ML predictive studies have been performed in youth internalizing disorders, these have to date considered <200 candidate predictors and focused largely on prevailing cases of depression, rather than new onset cases in adolescence, especially early adolescence. The latter are of considerable translational interest since understanding individual-level drivers of illness onset and obtaining better visibility into whether future onset can be reliably predicted using ML would potentially inform risk stratification strategies. Extant work is also highly heterogenous with respect to which candidate predictors (input features) are considered. In particular, some studies use only psychosocial features and some only neuroimaging features, while a few have incorporated both types. Concomitantly, performance has been variable, with accuracy ranging over ~50-90% but the achievement of robust precision (positive predictive value) - an important metric for translational relevance - typically proving more elusive. Moreover, since obtaining physiologic measures such as neuroimaging metrics is complex and uncommon in clinical practice, it is relevant to understand whether they improve individual-level case prediction. Finally, few studies have constructed predictive models of anxiety or somatic problems in youth using ML classifiers or applied a consistent analytic architecture across the three major categories of internalizing disorders simultaneously in the same population and data to enable direct comparisons and determine the specificity of predictive models to different internalizing disorders.

Beyond empirical findings, developmental and ecological frameworks also suggest that a variety of personal and environmental factors such as family context and self-regulatory processes are influential in the emergence of internalizing disorders. Bronfenbrenner's bioecological model emphasizes the broader influence of family and environment on child development, while Cicchetti and Rogosch's developmental psychopathology framework highlights how multiple interacting risks can lead to similar internalizing outcomes (26, 27). Transactional perspectives likewise point to the reciprocal shaping of child and parent behavior across developmental stages (28). Also, sleep has been conceptualized as a particularly important self-regulatory domain, with sleep disturbances proposed as a transdiagnostic risk factor for later internalizing symptoms (29–31).

In the present study, we aim to build on prior work by predicting cases of depression, anxiety and SSD in early adolescence (9–12 yrs) using deep learning guided by a large-scale AI optimization process. Specifically, we aimed to a) identify and rank the most important predictors after analyzing thousands of

multidomain candidate predictors; b) provide individual-level predictions of future, new onset cases at 11–12 yrs in comparison to all prevailing cases at the same age and 9–10 yrs; c) determine the incremental value of using multidomain predictors vs neural-only modeling; and d) examine the relationship between predictor importance and accuracy. Applying a common analytic architecture to data from the ABCD cohort, we first constructed multimodal predictive models by analyzing 5,810 candidate predictors spanning demographics; developmental and medical history; white and gray matter brain structure, neural function (cortical and subcortical connectivity, 3 tasks); brain volumetrics; physiologic function (e.g. sleep, hormone levels, pubertal stage, physical function); cognitive and academic performance; social and cultural environment (e.g. parents, friends, bullying); activities of everyday life (e.g. screen use, hobbies); living environment (e.g. crime, pollution, educational and food availability) and substance use. Subsequently, we recapitulated all analytic procedures using multiple types of neural candidate predictors.

To make these case classifications, we used deep learning with artificial neural networks, which incorporates non-linear relationships among predictors and is resistant to multicollinearity. While artificial neural networks offer powerful predictive capability, their application to translational aims can be limited by the relative difficulty of tuning these models (setting hyperparameters that control learning) and their tendency to act as 'black box' estimators where the features used to make predictions are not interpretable and their relative importance is difficult to determine. We enhanced deep learning performance with Integrated Evolutionary Learning (IEL), an AI-based form of computational intelligence, to jointly optimize across the hyperparameters and learn the most important final predictors and render explainable predictions. IEL is a genetic algorithm which instantiates the principles of natural selection in computer code, typically performing ~40,000 model fits during training before testing final, optimized models in a holdout, unseen data partition. All results presented are from testing for generalization in this holdout, unseen data.

Materials and methods

Terminology and definitions

Terms used in quantitative analysis may be shared among different fields with variant meanings. Here, we use ML conventions throughout (32–34). 'Prediction' means predicting the quantitative value of a target variable by analyzing patterns in input data. We refer to the set of all input data as containing 'features' or 'candidate predictors' and those identified in final, optimized models (presented in RESULTS) as 'final predictors'. The set of observations used to train and validate models is referred to as the 'training set' and the unseen holdout set of observations is termed the 'test set'. We use 'generalizability' to refer to the ability of a trained model to adapt to new, previously unseen data drawn from the same distribution i.e. model fit in the test set. 'Precision' refers to the fraction of positive predictions that were correct; 'Recall' to the proportion of true positives that were correctly predicted; and 'Accuracy' to the number of correct predictions as a fraction of total predictions. Receiver Operating Characteristic curves (ROC Curves) are provided that quantify classification performance at different classification thresholds plotting true positive versus false positive rates, where the Area Under the Curve (AUROC) is defined as the two-dimensional area under the ROC curve from (0,0) to (1,1).

Data and data collection in the ABCD study

Data used in the present study comes from the ABCD study, an epidemiologically informed prospective cohort study that is the largest study of brain development and child health conducted in the United States to date. ABCD recruited 11,880 children (52% male; 48% female) at ages 9–10 years (108–120 months) via 21 sites across the United States and will follow this cohort until age 19-20. The cohort is oversampled for twin pairs (n = 800) and non-twin siblings from the same family may also be enrolled. A wide variety of information is collected about participants. This data has been made available to qualified researchers at no cost by the NIH since 2018 and is released periodically. Currently, it may be obtained from the NBDC Data Hub (https://www.nbdc-datahub.org/). This study uses data from release 4.0, which includes data up to the 42-month follow-up date. A full explanation of recruitment procedures, the participant sample and overall design of the ABCD study may be found in Jernigan et al; Garavan et al; and Volkow et al. (35–37) This study has been reviewed and deemed not human subjects research by the University of Utah Institutional Review Board.

The phenotypic and substance abuse assessment protocol is covered in detail in Barch et al. and Lisdahl et al, respectively (38, 39). In brief, phenotypic assessments of physical and mental health, substance use, neurocognition and culture and environment are performed for youth and their parents and biospecimen collection for DNA, pubertal hormone levels, substance use metabolites (hair) and substance and environmental toxin exposure (baby teeth) are collected from youth at 9–10 yrs. A summary description of assessments performed and environmental and school-related variables derived from geocoding at age 9–10 yrs surveyed in the present study may be inspected in Supplementary Table S1.

Brain imaging is collected at 9–10 yrs and every two years thereafter and incorporates optimized 3D T1; 3D T2; Diffusion Tensor Imaging; Resting state functional MRI (rsfMRI); and 3 task MRI (tfMRI) protocols that are harmonized to be compatible across acquisition sites. The tfMRI protocol comprises the Monetary Incentive Delay (MID) and Stop Signal (SST) tasks and an emotional version of the n-back task which collectively measure reward processing, motivation, impulsivity, impulse control, working memory and emotion regulation. The ABCD study provides fully-processed metrics from each of these imaging types. Full details of the neuroimaging protocol may be inspected in Casey et al. and the pre-processing and analytic pipeline used to generate neural metrics in Hagler et al. (40, 41) The present study

TABLE 1 Demographic characteristics of participant sampled at age 9–10 years.

Characteristic	Number	Percent
Sex		
Male	2,663	51.8%
Female	2,473	48.2
Gender Identity		
Male	2,659	51.8%
Female	2,466	48.0
Gender non-conforming	7	0.1
Don't know/didn't answer	4	0.1
Race		
Black/African American	824	16.0%
Asian	373	7.3
White	4,069	79.2
Native American/Alaska Native	213	4.1
Other/don't know/didn't answer	390	7.6
Ethnicity		
Hispanic/Latino/Latinx	1,035	20.2%
Non-Hispanic	4,042	78.7
Not indicated	59	1.1

TABLE 2 Physiologic and cognitive characteristics of participant sample at age 9–10 years.

Characteristic	Range	Mean	Median
Age in months	107.0-132.0	120.0	120.0
Pubertal Development Stage	1-5	1.7	1.0
Height (inches)	36.6-74.0	55.4	55.5
Weight (pounds)	11.0-255.0	82.3	77.0
Waist Circumference (cm)	17.0-61.0	26.4	25.5
Handedness			
Writing	-100.0-100.0	76.5	100.0
Throwing	-100.0-100.0	67.1	100.0
Spoon	-100.0-100.0	70.2	100.0
Vocabulary	51.0-208.0	109.2	109.0
Attention and Inhibition	65.0-171.0	96.4	97.0
Working Memory	55.0-194.0	102.1	103.0
Executive Function	68.0-181.0	98.0	94.0
Processing Speed	20.0-185.0	95.2	95.0

Characteristics of the study sample at 9–10 yrs. Pubertal development is measured with the Pubertal Development Scale (adapted from the Petersen scale) in a sex-specific manner. Height is measured twice with the average of these values presented. We note a range of 11.0-255.0 pounds for weight which is the range present in the original ABCD data. Handedness is assessed with the Edinburgh Handedness Inventory. Cognitive metrics are assessed with the NIH Toolbox and are all age-corrected scores. Vocabulary is measured with the Picture Vocabulary Test; Attention and inhibition with the Flanker Inhibitory Control & Attention Test; Working Memory with the List Sorting Working Memory Test; Executive Function with the Dimensional Change Card Sort Test; and Processing Speed with the Pattern Comparison Processing Speed Test.
The resulting group of 5,136 participants was then randomly partitioned into a training set comprising 70% of the sample (n = 3,595) and a holdout, unseen test set comprising 30% of the sample (n = 1,541, Figure 1). This partitioning was performed prior to pre-processing either features or predictive target to minimize bias.

uses all available processed metrics that have passed quality control from the diffusion fullshell; cortical and subcortical Gordon correlations (derived from rsfMRI); structural; volumetric; and all three tasks as well as corresponding head motion statistics for each modality. For certain modalities such as rsfMRI, multiple scans were attempted or completed. In such cases we use variables from the first scan.

Study inclusion criteria and sample partitioning for machine learning

Inclusion criteria for the present study were a) participants enrolled in the study at baseline who were still enrolled at 2-year follow-up (n = 8,084) who had b) complete data passing quality control available for all neural metric types (n = 6,178) and were c) youth participants unrelated to any other youth participant in the study (n = 5,136). If a youth had a twin or other sibling(s) present in the cohort, we selected the older or oldest sibling for inclusion in our study. We present characteristics of the study sample at 9–10 yrs since these participants correspond to the input data used to make predictions. Demographic characteristics of this sample at age 9–10 yrs are presented in Table 1.

Sex refers to sex assigned at birth on the original birth certificate. Gender refers to the youth's gender identification. Race and ethnicity refer to the parents' view of youth's race or ethnicity. More than one race identification may be selected and therefore percentages sum to >100%.

Physiologic and cognitive characteristics of the participant sample at 9–10 yrs may be viewed in Table 2.

Steps in the formation of the study sample used to construct predictive models of depression, anxiety and somatic symptom disorder are shown. After exclusion criteria are applied, the sample was randomly partitioned into training and test sets followed by separate pre-processing of targets and features. Subsequently, samples for each experiment were formed as described in Preparation of predictive targets and Construction of participant case samples for internalizing disorders and controls.

Preparation of predictive targets

The present study uses predictive targets of depression, anxiety and somatic problems derived from the Child Behavior Checklist for youth ages 4–18 years (CBCL) called the 'ABCD Parent Child Behavior Checklist Scores Aseba (CBCL) in the ABCD study. The CBCL is a standardized instrument in widespread clinical and research use for the assessment of mental and emotional well-being in youth. It forms part of the Achenbach System of

Empirically Based Assessment (ASEBA) "designed to facilitate assessment, intervention planning and outcome evaluation among school, mental health, medical and social service practitioners who deal with maladaptive behavior in children, adolescents and young adults." (42) During assessment with the CBCL, parents rate their child on a 0-1–2 scale on 118 specific problem items such as "Unhappy, sad or depressed" or "Acts too young for age" for the prior 6 months. The answers to these questions are aggregated into raw, T and percentile scores for 8 syndrome subscales (Anxiety, Somatic Problems, Depression, Social Problems, Thought Problems, Attention Problems, Rule Breaking and Aggressive Behavior) derived from principal components analysis of data from 4455 children referred for mental health services. The CBCL is normed in a sex/gender-specific manner on a U.S. nationally representative sample of 2368 youth ages 4–18 that takes into account differences in problem scores for "males versus females". It exhibits excellent test-retest reliability of 0.82-0.96 for the syndrome scales with an average r of 0.89 across all scales. Content and criterion validity is strong with referred versus non-referred children scoring higher on 113/188 problem items and significantly higher on all problem scales, respectively.

To form binary classification targets for prediction, we thresholded CBCL subscale T scores for Depression, Anxiety and Somatic problems using cutpoints established by ASEBA for clinical practice. Specifically, a T score of 65-69 (95^{th} to 98^{th} percentile) is considered in the 'borderline clinical' range, and scores of ≥70 are considered in the 'clinical range.' Accordingly, we discretized T scores for each of the 3 subscales under consideration by deeming every individual with a T score ≥ 65 as a 'case' [1] and every individual with a score <65 as a 'not case' [0]. This process was performed separately for CBCL scores at baseline and 2-year follow-up in the training and test sets.

Construction of participant case samples for internalizing disorders and controls

To test our hypotheses, we formed 3 different participant samples for each of the internalizing disorders in the training and test sets, respectively (Figure 1). The first sample contained cases of depression, anxiety and SSD as defined in Preparation of predictive targets at baseline assessment, when youth were 9–10 years of age. The second sample contained cases of depression, anxiety and SSD at 2-year follow-up, when youth were 11–12 years of age. Finally, the third sample contained only new onset cases of depression, anxiety and SSD at 2-year follow-up. A new onset case was defined as a youth who met criteria for depression, anxiety or SSD following the ASEBA threshold in the CBCL who did not meet criteria for the disorder in question at baseline assessment. In all samples, we constructed a balanced sample of controls matched for age and sex/gender selected from the eligible study population (see: Study inclusion criteria and sample partitioning for machine learning above) from youth with the lowest possible scores on the relevant syndrome scale. No sample in the training sets was <200 participants, a recommended threshold for robust ML analyses.

Preparation of candidate predictors (input features)

The feature set in the present study comprises the majority of available phenotypic and environmental variables derived from baseline assessment at 9–10 years of age (including data collection site) and all available neural metrics (including head motion statistics) with the exception of temporal variance measures. For continuous phenotypic features where subscale or total scores for assessments were available, these were used. For example, subscale scores for different types of sleep-related disorders from the larger Munich Chronotype Questionnaire. Any metrics or instruments that directly quantified mental health symptoms were excluded since we aimed to predict cases of mental illness without using symptoms. For example, the Youth 7UP Mania scale. The feature set was then partitioned into training and test sets that conformed with the partitions detailed above in Formation of the study participant sample for internalizing disorders in Figure 1. Pre-processing of phenotypic and environmental features was subsequently performed separately in the training and test sets. First, features with >35% missing values were discarded. This threshold was used since prior research shows that good results may be obtained with ML methods with imputation up to 50% missing data (43). Also, it was selected pragmatically to balance the retention of potentially informative features against the risk of excessive imputation. Recent works have also suggested a slightly more lenient threshold (e.g., 40%) to exclude variables containing many missing values from analysis (44, 45). Nominal variables were one-hot encoded to transform them into binary variables. Continuous variables were then winsorized to [mean +/- 3] standard deviations to remove outliers, whereas ordinal variables were winsorized according to the bounds defined in the provided data dictionary. Then, all features were scaled in the interval [0,1] with the minimum-maximum normalization. Missing values were imputed using non-negative matrix factorization (NNMF). NNMF is a mathematically-proven imputation method that minimizes the cost function of missing data rather than assuming zero values. It can effectively reconstruct missing values in high-dimensional datasets by leveraging latent structure. It is effective at capturing both global and local structures in the data and has been demonstrated to perform well regardless of the underlying pattern of missingness (46–48). Supplementary Table S2 shows the number and percentage of observations which were trimmed and filled with NNMF for the training and test sets, respectively.

Further, a sensitivity analysis was performed to compare the predictive performance of NNMF with the Multiple Imputation by Chained Equations (MICE), which is a common alternative imputation method modeling each variable conditionally on the others and imputes missing values through iterative regression. Results of the sensitivity analysis (included in the Supplementary Table S4) demonstrate that test accuracy (when using our proposed algorithm described in Integrated Evolutionary Learning for deep learning optimization below) is generally consistent across the two imputation methods, with NNMF tending to slightly outperform MICE for most targets and case definitions (i.e., an average of 2.4%

10.3389/fpsyt.2025.1487894

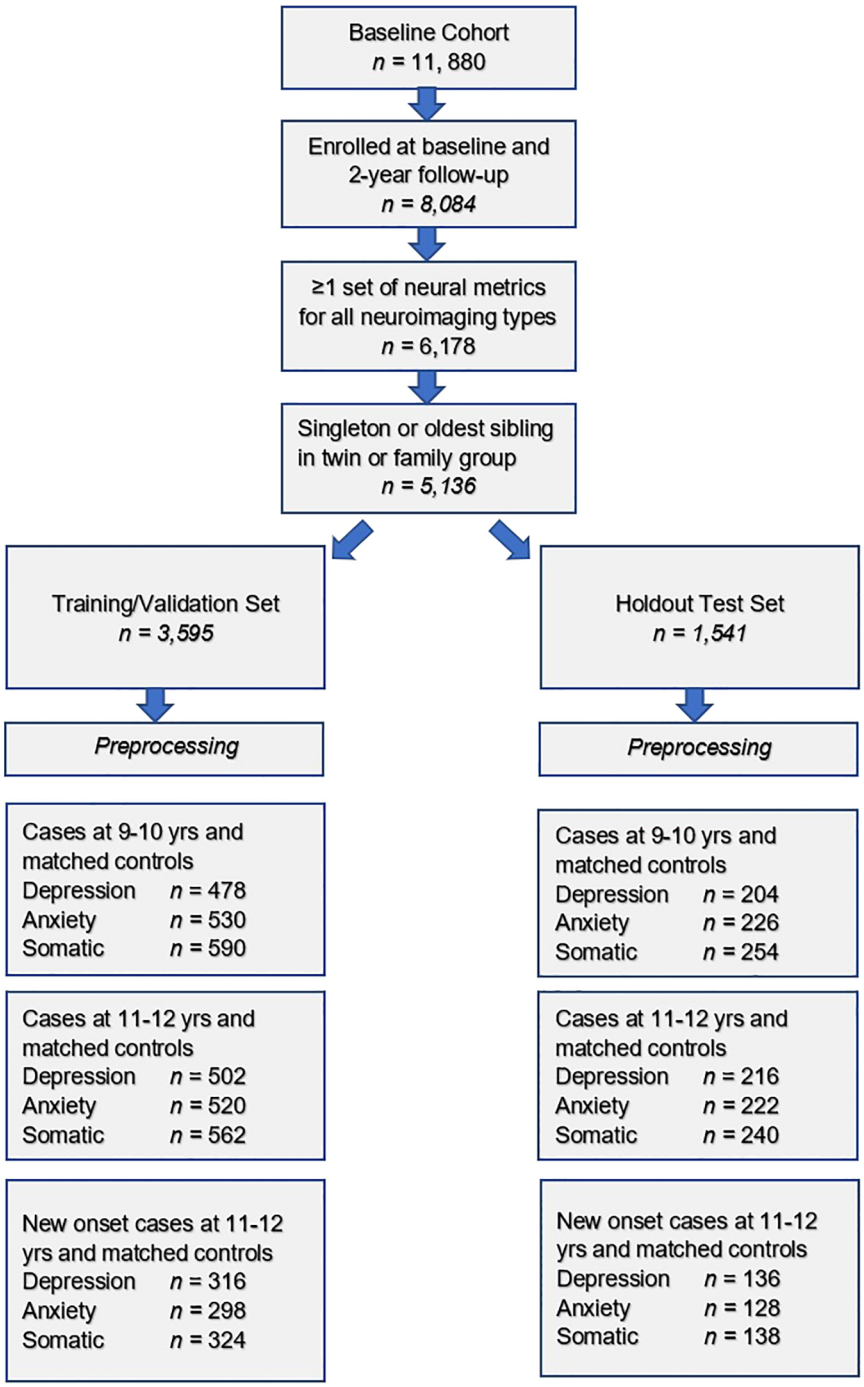

FIGURE 1
Formation of the study participant sample for internalizing disorders.

improvement in accuracy). In light of these findings, NNMF is selected to be the default data imputation method adopted in this work, as it provided equal or better performance in most scenarios while maintaining computational efficiency for high-dimensional datasets.

After imputation with NNMF, any variables originating from phenotypic assessments lacking summary scores were reduced to a summary metric using feature agglomeration to produce a final set of (n=804) phenotypic and environmental features. Neural metrics (n = 5,006) were processed and underwent quality control by the ABCD study team and were therefore not pre-processed with the exception of scaling, again performed separately in the training and test partitions. There were no missing neural features. The final combined feature set including neural, phenotypic, environmental, head motion and site features comprised 5,810 features.

Overview of predictive analytic pipeline

We used deep learning with artificial neural networks (AdamW optimizer) to predict cases of depression, anxiety and somatic problems in early adolescence in three scenarios: at 9–10 years of age, at 11–12 years of age and in new onset cases at 11–12 years of age. Deep learning models were implemented with *k*-fold cross-validation and trained by an AI meta-learning algorithm that jointly performed feature selection and optimized across the hyperparameters in an automated manner, pursuing ~40,000 model fits for each experiment. Model training was terminated based on the AUROC. Subsequently, final optimized models were tested for their ability to generalize in the holdout, unseen test set and performance statistics of AUROC, accuracy, precision and recall, and ROC curves are reported for the best-performing models. We also report the relative importance of final predictors to making case predictions quantified with two techniques: Shapley Additive Explanations (SHAP) and permutation using the eli5 algorithm. Detailed explanations of these methods are provided below. Code for the predictive analytics may be accessed at the de Lacy Laboratory GitHub: https://github.com/delacylab/integrated_evolutionary_learning.

Coarse feature selection

Prior to beginning model training, we performed coarse feature selection for each of the nine experiments i.e. 3 targets of depression, anxiety and SSD each in 3 participant samples of 9–10 yrs; 11–12 yrs and new onset cases at 11–12 yrs. The purpose of this process was to quantify, for each sample, which of the 5,810 features exhibited a non-zero relationship with the target in order to reduce the number of features entering the deep learning pipeline in a principled manner. First, a simple filtering process was performed in which χ^2 (categorical features) and ANOVA (continuous features) statistics and mutual information metric (all features) were computed to quantify the relationship between all features and the target, where the target (depression, anxiety, SSD) was represented by a categorical vector in [0,1]. Any feature with a non-zero relationship (either positive or negative) with the target was retained. While we acknowledge the potential risk that this univariate filtering cannot fully capture complex interactions across the features, it is practically advantageous to reduce the dimensionality of the feature space before the execution of additional filtering procedures that require high computational complexity in face of large feature sets.

Subsequently, feature selection was performed on these filtered feature subsets using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm. LASSO is a popular regularization technique based in linear regression that efficiently selects a reduced set of features by forcing certain regression coefficients to zero. The LASSO algorithm has a hyperparameter (commonly called the α) that instantiates the amount of penalization (shrinkage) that will be imposed on the features. We implemented the LASSO with our AI meta-learning algorithm Integrated Evolutionary Learning to tune the α hyperparameter in the same manner as described below in Integrated Evolutionary Learning for deep learning optimization.

Notice that the LASSO algorithm may lead to biased feature selection results when multicollinearity exists in the pre-selected feature set or features were related to the target in a non-linear manner. To alleviate these potential problems, Boruta (49) was selected as a complementary feature selection method. Boruta is an ensemble-based method designed to capture all relevant features using random forest modeling and has been shown to be less sensitive to multicollinearity (50). A sensitivity analysis comparing the LASSO only, the Boruta only, and the LASSO combined with Boruta methods was performed. Since the feature set selected by Boruta tends to be small but non-overlapping with that selected by LASSO, we focused on the comparison between LASSO and the LASSO combined with Boruta methods. Results from the sensitivity analysis, included in Supplementary Table S5, demonstrated that the combined method yields 1-10% improvement in predictive accuracy (when using our proposed Integrated Evolutionary Learning models explained in the next subsection) compared to the LASSO only method in all targets and ages of case determination. In light of this improvement, this combined method, which selects the union of features retained by both LASSO and Boruta, is taken as the default feature selection configuration in this study unless specified otherwise.

The number of features retained for each of the 9 experiments after each step in the coarse feature selection process may be examined in Table 3. Specific features selected by the LASSO combined with Boruta and the resulting feature importance scores (univariate coefficients in LASSO and Boruta importance scores) between each of these features and the target vectors (depression, anxiety, somatic problems) for each participant sample (9–10 yrs; 11–12 yrs and new onset cases at 11–12 yrs) may be viewed in Supplementary Table S3a-i. Each feature set selected by the LASSO combined with Boruta then entered the deep learning pipeline.

The total baseline set of 5,810 features was reduced via coarse feature selection in a two-step process of filtering followed by regularization with the LASSO algorithm combined with the Boruta algorithm. This table displays the number of remaining features after each step for each target (depression, anxiety and somatic problems) and participant sample (at age 9–10 years, at age 11–12 years and for new onset cases at age 11–12 years). Detailed tables showing the univariate coefficients between each feature selected by the LASSO and the target vectors for each case sample and controls may be viewed in Supplementary Table S3a-i.

Deep learning with artificial neural networks

We used deep learning to predict cases of depression, anxiety and somatic problems in each participant sample (at ages 9-10, ages 11–12 and for new onset cases only at ages 11–12 years). In order to determine the relative ability of features to predict future cases of internalizing disorders, features collected at baseline assessment

10.3389/fpsyt.2025.1487894

TABLE 3 Feature sets after coarse feature selection.

Condition and age of case determination	Number of features after filtering	Number of features after selection with LASSO combined with Boruta
Depression, age 9–10 years	5,783	140
Anxiety, age 9–10 years	5,782	152
Somatic problems, age 9–10 years	5,777	96
Depression, age 11–12 years	5,779	58
Anxiety, age 11–12 years	5,783	131
Somatic problems, age 11–12 years	5,786	207
Depression, new onset age 11–12 years	5,773	70
Anxiety, new onset age 11–12 years	5,767	97
Somatic problems, new onset age 11–12 years	5,764	128

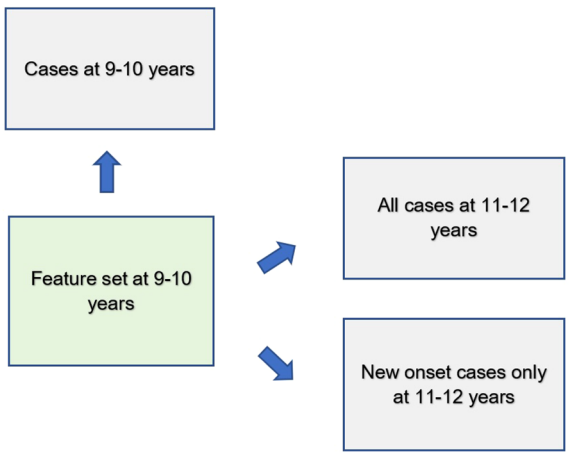

FIGURE 2
Analytic schema.

(ages 9–10 years) were used to predict cases present at ages 11–12 years. We also constructed similar models that restricted the cases at 11–12 years of age to only new onset cases, where the participant was not exhibiting clinical levels of symptoms at ages 9–10 years. Finally, to quantify any dropoff in predictive power over the two-year followup period, comparative models predicting cases at 9–10 years of age were also computed. Therefore, the feature set comprised only variables collected at 9–10 years of age in all analytic scenarios (Figure 2).

TABLE 4 Hyperparameter settings optimized with Integrated Evolutionary Learning.

Hyperparameters	Range	Mutation shift
Learning rate	0.00001-0.01	0.0001
Beta 1	0.9-0.999	0.001
Beta 2	0.9-0.999	0.001

Optimization across the hyperparameters of learning rate, Beta 1 and Beta 2 was conducted for deep learning with artificial neural networks within the ranges shown.

Features assessed at baseline (ages 9–10 years) were used to predict cases of depression, anxiety and somatic problems present contemporaneously as well as all cases 2 years in the future (ages 11–12 years) and only new onset cases at ages 11–12 years.

We trained artificial neural networks using the AdamW algorithm with 3 layers, 300 neurons per layer, early stopping (patience = 3, metric = validation loss) and the ReLU activation function. The last output layer contained a conventional softmax function. Learning parameters (Table 3) were tuned with IEL as detailed below. Deep learning models were encoded with PyTorch embedded in custom Python code (51).

Integrated Evolutionary Learning for optimization across hyperparameters and fine feature selection

Many ML algorithms have hyperparameters that control learning. Their settings require 'tuning' that can have a dramatic effect on performance. Typically, tuning is performed via 'rules of thumb' and ≤50 model fits are explored, introducing the possibility of bias and potentially limiting the solution space (52–54). To address this issue, we previously developed and here applied an AI technique called Integrated Evolutionary Learning (IEL) which can improve the performance of ML predictive algorithms in comparable tabular data by up to 20-25% versus the use of default model hyperparameters and conventional designs (55). IEL is a form of computational intelligence or metaheuristic based on an evolutionary algorithm that instantiates the concepts of biological evolutionary selection in computer code. It optimizes across the hyperparameters of the deep learning algorithm by adaptively breeding models over hundreds of learning generations by selecting for improvements in a fitness function (here, AUROC).

For each experiment, the deep learning algorithm was nested inside IEL, which initialized the first generation of 100 models with randomized hyperparameter values or 'chromosomes'. These hyperparameter settings (Table 4) were subsequently recombined, mutated or eliminated over successive generations. In recombination, 'parent' hyperparameters were arithmetically averaged to form 'children'. In mutation, hyperparameter settings were shifted with the range of possible values shown in Table 4. When these first 100 models were trained, the BIC was computed for each solution. Of the 80 best models, 40 were recombined by

averaging the hyperparameter setting after a pivot point at the midpoint to produce 20 'child' models. 20 were mutated to produce the same number of child models by shifting the requisite hyperparameter by the mutation shift value (Table 4). The remaining 20 were discarded. The next generation of models was then formed by adding 60 new models with randomized settings and adding these to the 40 child models retained from the initial generation. Thereafter, IEL continued to recombine, mutate and discard 100 models per generation in a similar fashion to minimize the BIC until the latter fitness function plateaued. With 100 models fitted per generation, IEL typically fits ~40,000 models per experiment over ~400 generations.

IEL jointly performs optimization across hyperparameter settings with automated feature selection and mitigate the risk of overfitting. For each experiment, IEL has available to it the set of features selected in the two-step feature selection process performed with filtering and the LASSO (Coarse feature selection, Supplementary Table S3). From each of these sets, a random number of features in the range [2-50] was set for each model in the initial generation of 100 models and specific features were randomly sampled from the set of available features. This iterative subset exploration reduces the risk that predictive performance hinges on a single subset or on correlated features retained by the initial selection. After computing the AUROC for each model, feature sets from the best-performing 60 models were individually allocated to the recombined and mutated child models. Other feature sets were discarded. As with hyperparameter tuning, this process was repeated for succeeding generations until the AUROC plateaued.

IEL implements recursive learning to facilitate computationally efficiency. After training until the AUROC plateaued, we determine the elbow of the fitness function plotted versus number of features and re-start learning with a warm start. The feature set available after this warm start is constrained to that subset of features, thresholded by their importance, corresponding to the fitness function elbow. Learning then proceeds by thresholding features available for learning at the original warm start feature importance + 2 standard deviations. In addition, the number of models per generation is reduced to 50 and 20 models are recombined and 10 models are mutated. Otherwise, training after the warm start uses the same principles as detailed above.

Cross validation

Deep learning models were fit within IEL using stratified *k*-fold cross validation i.e. every one of the 100 models in each learning generation within IEL was individually trained and validated using cross-validation in the training partition. IEL allows the number of features used to fit each model to differ within each model in every generation. Accordingly, *k* (the number of splits) was set as the nearest integer above [sample size/number of features]. Cross validation was implemented with the scikit-learn StratifiedKFold function.

Testing for generalization in holdout, unseen test data and performance measurement

After training was completed, optimized models generated by IEL were tested on the holdout, unseen test set for each sample and mental health condition by applying the requisite hyperparameter settings and selected features obtained from the 100 best-performing models in the training phase to the test set. The area under the receiver operating curve (AUROC), accuracy, precision, and recall were computed for test set models using standard Sci-Kit learn libraries and models with the best performance in each statistic selected for presentation as the final, optimized models. The threshold for prediction probability was 0.5 and receiver operating characteristic (ROC) curves are also provided for each experiment (Supplementary Figures 1, 2).

Feature importance determination

Shapley Additive Explanations (SHAP) values were computed to determine the relative importance of each feature to predicting cases of mental illness. SHAP is a game theoretic approach commonly used in ML to explain the output of any ML model including 'black box' estimators such as artificial neural networks and is considered resistant to multicollinearity (56). In this work, GradientSHAP, encoded in PyTorch (51) and Captum (57) packages in Python as a combined technique of Integrated Gradients (58) and SmoothGrad (59), is adopted as a fast approximation of SHAP values for gradient-based models.

To illustrate the model's decision-making process in clinically interpretable terms, a case study is provided for each target's onset cases at 11–12 years. In each case, the contribution from the key features to the predicted labels of two selected contrasting participants were studied in terms of GradientSHAP values. The visualized GradientSHAP values in Figures 3–5 serve as indicators of risk versus protective factors of the relevant mental illness.

Fairness subgroup performance analysis

While the ABCD dataset is one of the most demographically diverse pediatric neuro-developmental cohorts currently available, it may not fully represent all racial and socioeconomic groups within the U.S. population. Thus, even if a trained model has a decent predictive performance in general, it may not perform as well in each demographic subgroup. To robustly evaluate the model performance from a fairness perspective, we conducted a subgroup performance analysis on the held-out test set, stratified by race and annual family income level, in order to assess whether models perform consistently across subpopulations.

The stratification of race is conducted in terms of 3 racial subgroups: White, Black, and other/multiple races (Others). Annual family income level was stratified into 4 subgroups: below $15,999,

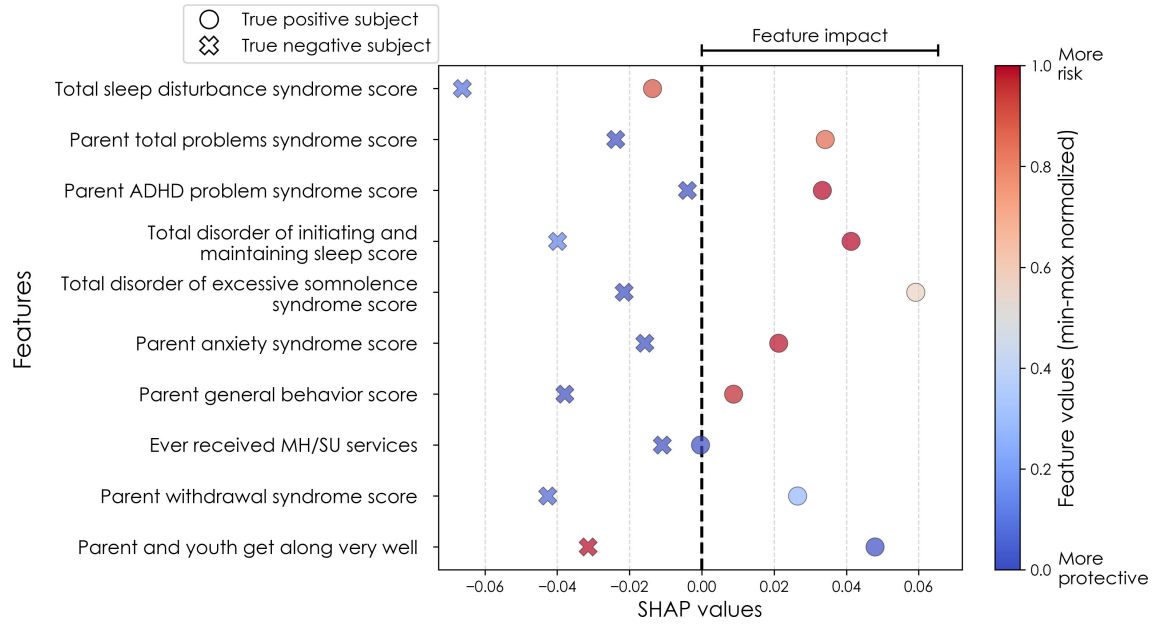

FIGURE 3
Case study of depression predictors in the multimodal predictive model of new onset at 11–12 years. The summary plot presents the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting depression with new onset at 11–12 yrs. The color gradient represents the scaled value of each feature where red = high and blue = low. Circle and cross marks correspond to the feature values of a pre-selected pair of (true) positive subject and (true) negative subject respectively. MH, mental health; SU, substance use.

$16,000-$34,999, $35,000-$74,999, and above $75,000). The fairness subgroup analysis considers only the onset cases of 11–12 years of each studied internalizing disorders, which are the main focus groups in this work. For each subgroup analysis, we evaluated the predictive performance in terms of accuracy, precision, recall, and AUROC to assess whether the predictive model systematically performs better or worse across subpopulations.

Baseline modeling comparison

To assess the predictive performance of the IEL modeling technique, we conducted a baseline modeling comparison using several traditional ML modeling methods, including logistic regression, random forest, and support vector machines (SVM). Thees models were trained on the same multimodal feature sets obtained after the LASSO combined with Boruta feature selection step (i.e., containing 58–207 features). This aims to compare whether the IEL modeling pipeline, which embeds an iterative feature inclusion step in its evolutionary process, can result in consistent predictive performance compared to the baseline benchmarking models while involving strictly fewer features. All baseline models were evaluated using the same train/test splits as our deep learning approach and encoded using the Python package sklearn (60) with their default runtime parameters. Student t-tests were performed to verify whether the predictive accuracy score of IEL was statistically significantly different from the scores of the three baseline modeling methods.

Summary of the preprocessing and modeling pipelines

To improve the transparency and reproducibility of our study, the data preprocessing and modeling pipelines are documented on our code-sharing space (https://github.com/delacylab/integrated_evolutionary_learning). Below we recap the overall pipelines for clarity.

- Variables with less than 35% of missing values were retained.
- The full samples were partitioned into a training set for model training and a held-out test set for model evaluation.
- For each of the training set and test set, variables were winsorized, scaled to the unit interval, and imputed.
- Coarse feature selection was performed to retain features that exhibited a non-zero relationship with the target in the training set.
- Feature selection combining LASSO and Boruta was performed to retain relevant features.
- Cross-validated deep learning models were trained within the IEL algorithm to perform hyperparameter optimization and fine feature inclusion.
- Identify the elbow of the fitness curve (obtained across the generations run in the IEL algorithm) to shrink the feature subset further.
- Re-train the IEL algorithm with a warm-started feature subset until the fitness score plateau.

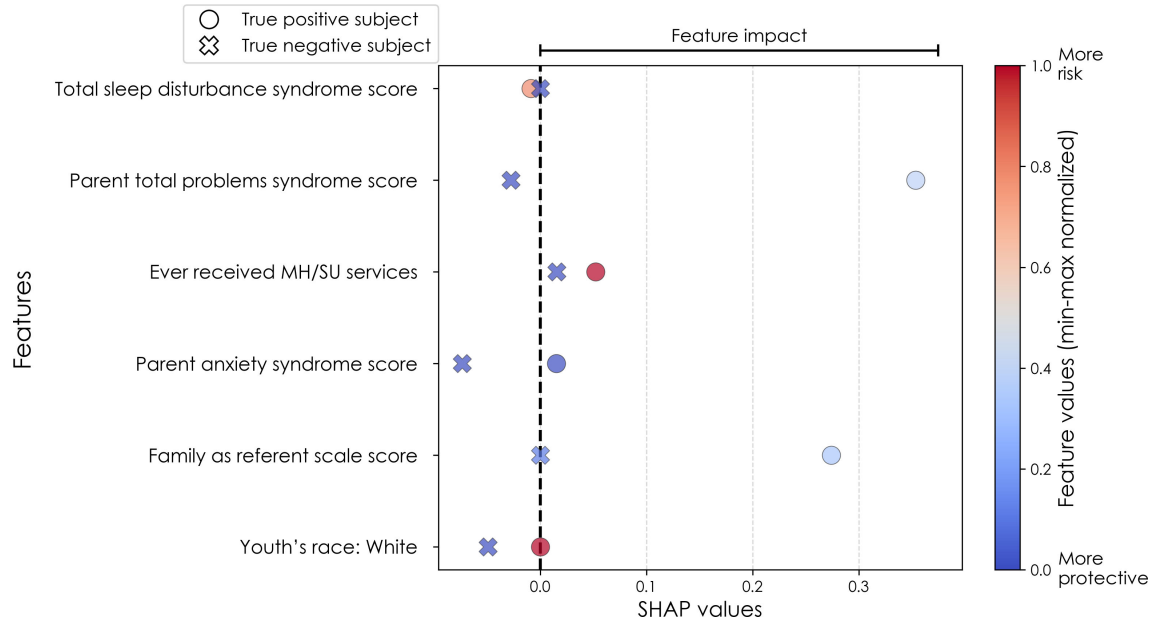

FIGURE 4
Case study of anxiety predictors in the multimodal predictive model of new onset at 11–12 years. The summary plot presents the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting depression with new onset at 11–12 yrs. The color gradient represents the scaled value of each feature where red = high and blue = low. Circle and cross marks correspond to the feature values of a pre-selected pair of (true) positive subject and (true) negative subject respectively. MH, mental health; SU, substance use; SST, Standard Stop Signal task; MID, Monetary Incentive Delay task; ROI, region of interest; FA, fractional anisotropy; LD, longitudinal diffusivity; WM, white matter; GM, gray matter.

- Evaluate the test data set using the model retrained with the hyperparameter settings and selected features of the best-fitting model identified in the last execution of IEL.

Results

Overview

All results are from testing the final model obtained after optimization with IEL for generalization in the holdout, unseen test dataset for each participant sample and experiment. For each condition (depression, anxiety, SSD) a parallel set of results is presented for each participant sample of new onset cases at 11–12 yrs; all prevailing cases at 9–10 yrs and all prevailing cases at 11–12 yrs. In all experiments only data collected at 9–10 yrs is input to deep learning to make predictions. Thus, results obtained for new onset and prevailing cases at 11–12 yrs represent predictions of future case status.

For each disorder and age group, results are presented for the metrics below for a) multimodal models constructed using all types of input features; and b) neural-only models.

- Performance statistics: accuracy, precision, recall and AUROC. ROC curves may be viewed in Supplementary Figures 1, 2.
- Final predictors ranked in order of importance by their group-level SHAP score (average absolute value across the participant sample) and the mean predictor importance for the requisite experiment.
- Individual-level final predictor importance (SHAP scores) across the participant sample. This summary plot is also used to determine the directionality of the relationship between the predictor and case status.

Results summary

Across all prediction scenarios, our models consistently achieved strong discriminative ability, particularly in multimodal settings where phenotypic, environmental, and neural predictors were included. In all cases, multimodal models outperformed neural-only models (see Tables 5–7), underscoring the central importance of psychosocial domains in early prediction of internalizing disorders. Predictive models are essential precursors of risk stratification tools, and we note that our models here achieved very strong positive class discrimination, with precision (positive predictive value) of 77-84%. Parental psychopathology and child sleep disturbances emerged as important cross-cutting predictors across outcomes (see Figures 6–8). Our proposed IEL modeling strategy achieved highly comparable performance with the traditional ML modeling methods with the primary benefit of improved model parsimony, another important feature when creating risk stratification precursor models. We note that results from our fairness subgroup analyses demonstrate visible disparities

in predictive performance across different racial groups and family income levels.

Depression

Deep learning optimized with IEL predicted depression in early adolescence with >80% accuracy and recall and ≥90% AUROC across all experiments (Table 5a), with precision of ~77-84%. Performance was slightly worse by a few percentage points in predicting new onset cases in the future (at 11–12 yrs) than either contemporaneous or all prevailing cases at 11–12 yrs. When each experiment was recapitulated using only neural candidate predictors, we found that final optimized predictive models displayed substantially lower performance (Table 5b) than those obtained with multimodal predictors with accuracy of ~52-56% and AUROC of ~56-60%, or some 27-39 percentage points lower than with multimodal predictors. Similar differentials were seen in precision and recall. In depression, multimodal models achieved somewhat better performance when predicting prevailing cases at 9-10yrs and11–12 vs new onset cases at 11–12 yrs. In neural-only models this was reversed, with a substantially stronger model obtained for new onset cases.

Performance statistics of accuracy, precision, recall and the AUROC are shown for the most accurate model obtained with deep learning optimized with Integrated Evolutionary Learning using a) multimodal features and b) only neural features. We used features obtained at 9–10 years of age to predict new onset cases of depression at 11–12 years of age as well as all prevailing contemporaneous cases (9–10 yrs) and all prevailing cases at 11–12 years of age. Corresponding ROC curves may be viewed in Supplementary Figures 1 and 2.

In interpreting multimodal models (Table 8), we found that parent problem behaviors were the most important predictors of early adolescent depression in each participant sample. Specific parental behavioral drivers of youth cases differed by age and case type. In new onset cases at 11–12 yrs, positive parent-youth relationships were the top predictor, followed by parent sleep disturbance, withdrawal traits, and excessive somnolence, while indicators of overall parental behavioral burden and prior mental health/substance use (MH/SU) services were also present. In all cases at 9–10 yrs, parent avoidant and somatic traits were most important, along with sleep disturbances and prosocial behaviors. In all cases at 11–12 yrs, parent behavioral problems and sleep disturances were most predictive, with positive parent-youth relationships and prosocial behaviors appearing among the top-ranked features. Group-level importances for multimodal model predictors (averaged across the participant sample) were in the range [0.05, 0.07] and the mean importance for each experiment in the range [0.018, 0.05].

Final predictors of cases of all prevailing cases of depression at ages 9–10 and 11–12 years as well as new onset cases only at 11–12 years of age are shown for the most accurate models obtained using deep learning optimized with IEL obtained with a) multimodal features and b) only neural features. Final predictors are ranked in order of importance where the relative importance of each predictor is computed with the Shapley Additive Explanations technique and presented here averaged across all participants in the sample. Features in blue indicate an inverse relationship with depression verified with the Shapley method. MH = mental health; SU = substance use; SST = Standard Stop Signal task; MID = Monetary Incentive Delay task; ROI = region of interest; FA = fractional anisotropy; LD = longitudinal diffusivity; WM = white matter; GM = gray matter.

Final predictors of new onset cases at 11–12 yrs obtained in neural-only models (Table 8b) were dominated by features derived from the Standard Stop Signal fMRI task, which measures response inhibition. Here, SST ROIs emphasized the left hemisphere. Specifically, SST responses in pars opercularis (Broca's area), left frontal pole, anterior cingulate and transverse temporal area. Certain structural metrics also appeared as final predictors of new onset cases. Specifically, white-gray contrast in the right superior temporal sulcus, right accumbens T1 intensity and white matter structural integrity in the left anterior cingulate.

Connectivity metrics in sensorimotor and cingulo-parietal brain networks were prominent in predicting contemporaneous prevailing cases of depression at 9-10yrs as were, again, metrics associated with the Stop Signal task – once again in Broca's area (pars triangularis) and frontal regions. The final predictive model for all prevailing future cases of depression at 11–12 yrs was parsimonious and included contrast differences in the right inferior temporal ROI in the Monetary Incentive Delay task, which measures approach and avoidance during reward processing, and correlation between the auditory functional connectivity neural network and the right accumbens. Group-level importances for neural-only model predictors were in the range [0.11,0.30] and the mean importance for each experiment in the range [0.14, 0.18]. As indicated by the color-coding of the feature values in Figures 6 and 9, feature values of the neural predictors generally have a smaller variance than the psychosocial predictors in the multimodal models.

Where Table 8 presents the importance of final predictors as summarized (mean absolute value) across the requisite experimental participant sample, we were also interested in predictor importance at the individual participant level. We computed and plotted individual-level SHAP values to understand both the dispersion of predictor importances across individuals and the directionality of the relationship between final predictors and clinical case status (Figure 6). In SHAP summary plots, each data point represents an individual participant and the colorization reflects the original value of the predictor as an input feature. Thus, discrete-valued features appear as red or blue, whereas a continuous feature appears as a color gradient from low to high. The directionality of the relationship between predictors and depression case status obtained in these plots was further compared with coefficients obtained during LASSO regression for Coarse Feature Selection (Supplementary Table S3) and found to be in agreement.

Figure 6 reveals that individual-level importance of final predictors in early adolescent depression are typically widely

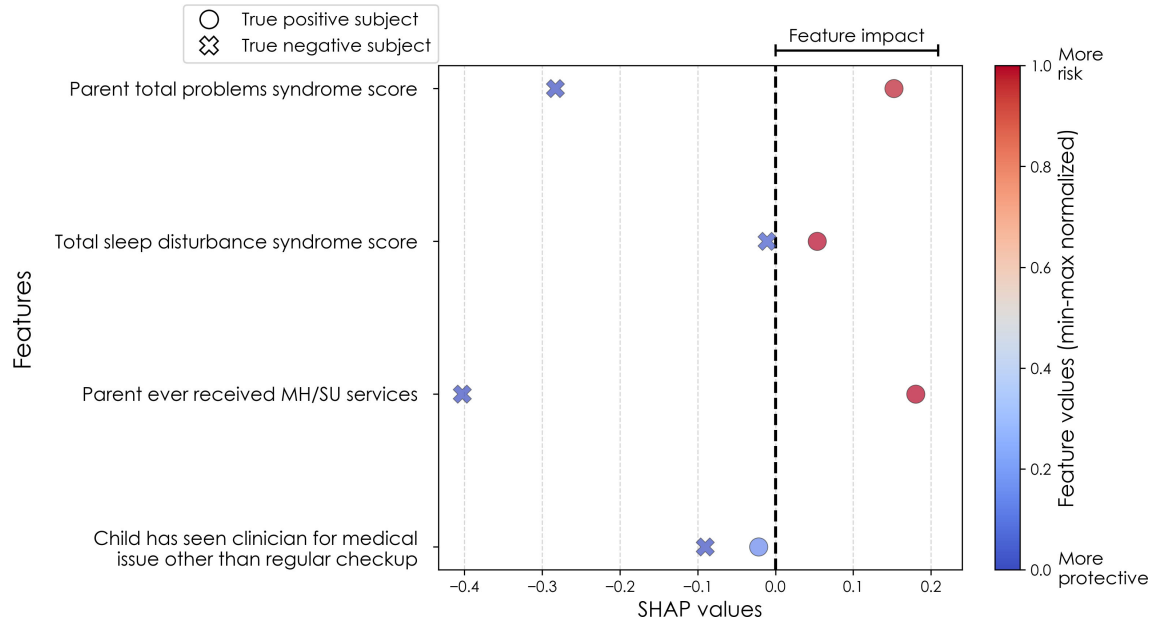

FIGURE 5
Case study of predictors of somatic disorder in the multimodal predictive model of new onset at 11–12 years. The summary plot presents the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting depression with new onset at 11–12 yrs. The color gradient represents the scaled value of each feature where red = high and blue = low. Circle and cross marks correspond to the feature values of a pre-selected pair of (true) positive subject and (true) negative subject respectively. MH, mental health; SU, substance use.

TABLE 5 Performance of deep learning optimized with integrated evolutionary learning in predicting cases of depression using multimodal and neural-only feature types.

a				
Age of case determination	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	83.1	86.9	77.9	93.3
All cases at age 9–10 years	91.7	92.9	90.2	96.1
All cases at age 11–12 years	87.0	89.2	84.3	94.5
b				
Age of case determination	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	55.9	54.5	70.6	60.4
All cases at age 9–10 years	52.5	52.4	53.9	56.3
All cases at age 11–12 years	56.5	57.1	51.9	57.3

dispersed. For example, when predicting new onset cases of depression at 11–12 yrs, the leading predictor of parent externalizing traits has a large range of ~[-0.4,0.6] across individual participants. Further, dispersion is typically greater for the more important predictors. Overall, these plots also indicate that all final predictors obtained have a positive relationship with depression case status, with the exception of secondary caregiver acceptance and prosocial behaviors in predicting new onset cases (see also Table 8). We also computed individual-level importances of final predictors for neural-only experiments (Figure 9). Here, the dispersion of individual-level predictor importances across participants were consistently smaller in neural-only versus multimodal prediction of early adolescent depression. In addition, Figure 3 visualized the individual-level predictor importances of two selected contrasting participants studied at the new onset cases at 11–12 yrs, which further differentiate clearly the protective

10.3389/fpsyt.2025.1487894

TABLE 6 Performance of deep learning optimized with IEL in predicting cases of anxiety.

a				
Age of case determination	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	78.9	76.1	84.4	85.6
All cases at age 9–10 years	94.2	97.2	91.2	98.5
All cases at age 11–12 years	83.8	83.8	83.8	91.9
b				
Neural data type	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	51.6	51.9	43.8	56.6
All cases at age 9–10 years	53.5	54.1	46.9	55.3
All cases at age 11–12 years	50.0	50.0	53.2	53.1

TABLE 7 Performance of deep learning optimized with integrated evolutionary learning in predicting cases of somatic symptom disorder.

a				
Age of case determination	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	82.6	83.6	81.2	90.4
All cases at age 9–10 years	90.9	91.9	89.8	95.2
All cases at age 11–12 years	82.5	86.1	77.5	91.4
b				
Neural data type	Accuracy (%)	Precision (%)	Recall (%)	AUC
New onset at age 11–12 years	54.3	54.8	49.3	52.2
All cases at age 9–10 years	46.1	46.2	48.0	44.7
All cases at age 11–12 years	53.8	53.1	63.3	56.2

factors (e.g., parents and youth getting along very well) from the risk factors (e.g., parents' syndrome scores of multiple mental illnesses).

Anxiety

Deep learning optimized with IEL performed very well in predicting both new onset and prevailing cases of anxiety in early adolescence. In anxiety, ~79% accuracy and ~86% AUROC was achieved in predicting new onset cases versus ~84% accuracy and ~92% AUROC in predicting prevailing cases using data obtained at 9–10 yrs to predict cases at the future time point of 11–12 yrs. The best overall performance was observed using data at 9–10 yrs to predict contemporaneous prevailing cases, with ~94% accuracy and nearly 100% AUROC achieved (Table 6a). Similar to depression, neural-only models did not perform as well as multimodal models in predicting anxiety cases, being ~24-40% less accurate. Best performance was obtained when predicting all cases of anxiety at 11–12 yrs, where the final, optimized neural-only model achieved 54% accuracy and ~55% AUROC. Neural-only predictive models of all prevailing cases at 9–10 yrs and 11–12 yrs also showed inferior performance with accuracy of ~52 and ~50% and AUROC of 57 and 53% respectively (Table 6b).

Performance statistics of accuracy, precision, recall and the AUROC are shown for the most accurate model obtained with deep learning optimized with Integrated Evolutionary Learning using a) multimodal features and b) only neural features. We used features obtained at 9–10 years of age to predict new onset cases of anxiety at 11–12 years of age as well as all prevailing contemporaneous cases (9–10 yrs) and all prevailing cases at 11–12 years of age. Corresponding ROC curves may be viewed in Supplementary Figures 1 and 2.

In anxiety, new onset cases were predicted with a relatively complex final model comprising 5 predictors (Table 9a). Here, the most important predictor was the parent's total burden of behavioral problems, followed by parent anxiety traits, whether the youth had ever received mental health or substance use (MH/SU) services, the youth's race (White), and the family's referent scale score. Total sleep disturbance also appeared but with lower importance.

We detected overlap between the final predictors of new onset cases of anxiety at 11–12 yrs and those which predicted prevailing

cases at 9–10 and 11–12 yrs. At 9–10 yrs, parent depressive and total behavioral problem scores were the top predictors, with sleep disturbances and parent–youth relationship quality also appearing. For all prevailing cases at 11–12 yrs, sleep disturbances, parent anxiety and aggressive traits, and the mother's history of clinical treatment were prominent, along with overall parent behavioral burden. Group-level importances for multimodal model predictors were in the range [0.03, 0.18] and the mean importance for each experiment in the range [0.05, 0.09].

Final predictors of cases of all prevailing cases of anxiety at ages 9–10 and 11–12 years as well as new onset cases only at 11–12 years of age are shown for the most accurate models obtained using deep learning optimized with IEL obtained with a) multimodal features and b) only neural features. Final predictors are ranked in order of importance where the relative importance of each predictor is computed with the Shapley Additive Explanations technique and presented here averaged across all participants in the sample. Features in blue indicate an inverse relationship with depression verified with the Shapley method. MH = mental health; SU = substance use; SST = Standard Stop Signal task; MID = Monetary Incentive Delay task; ROI = region of interest; FA = fractional anisotropy; LD = longitudinal diffusivity; WM = white matter; GM = gray matter.

In neural-only models predicting new onset anxiety cases, features from the MID fMRI task and structural metrics predominated (Table 9b). Important final predictors were cortical depth in the left hemisphere and right pars triangularis (Broca's area) as well as MID anticipation in the latter. Final, optimized models predicting prevailing cases at 9–10 years emphasized measures of brain function including SST and nBack task metrics in the right postcentral and left temporal, as well as multiple connectivity measures including the ventral attention, default mode, sensorimotor and cingulo-opercular networks. At 11–12 yrs, results again emphasized connectivity metrics among control and sensorimotor networks as well as structural measures. Group-level importances for neural-only model predictors were in the range [0.09, 0.16] and the mean importance for each experiment in the range [0.12, 0.13]. Similar to the case of depression, the feature values (as visualized in Figures 7, 10) of the neural predictors had a relatively smaller variance than the psychosocial predictors in the multimodal models.

To probe the dispersion of predictor importances at the individual level, we again developed summary plots of individual-level importances (Figures 7, 10). Similarly to depression, we observed relatively more widely dispersed individual-level importances over the participant sample in multimodal vs neural-only models, and the trend for wider dispersion of predictor importance in the more important final predictors. The directionality of the relationship between predictors and depression case status obtained in these plots was further compared with coefficients obtained during LASSO regression for Coarse Feature Selection (Supplementary Table S3) and found to be in agreement.

Individual-level predictor importances for the best-performing mixed-type neural models of anxiety again showed reduced dispersion across the participant group (Figure 10) when compared with multimodal models (Figure 7). The widest dispersion was observed when predicting new onset cases of anxiety. Figure 4 presents the individual-level predictor importances of two selected contrasting participants studied at the new onset cases at 11–12 yrs. Unlike the case study for depression, the current case study provides a less obvious disparity, which can potentially be explained by the relatively weaker predictive performance of the model predicting anxiety onset at 11–12 yrs.

Somatic symptom disorder

Deep learning optimized with IEL performed well using multimodal data in predicting both new onset and prevailing cases of SSD in early adolescence. Here, ~83% accuracy and ~90% AUROC was achieved in predicting future, new onset cases at 11–12 yrs with data obtained at 9–10 yrs. The best overall performance was observed using data at 9–10 yrs to predict contemporaneous prevailing cases, with ~91% accuracy and ~95% AUROC. Predictive performance of all prevailing cases at 11–12 yrs using data from 9-10yrs was comparable to new onset predictions, with accuracy of ~83% and AUROC of ~91% (Table 7a). As with depression and anxiety, neural-only models did not perform as well as multimodal models (Table 7b), being ~29-45% less accurate. The best performance was seen in predicting new onset cases at 11–12 yrs with accuracy of ~54% and AUROC of ~52% and all prevailing cases at 11-12 yrs with accuracy of ~54% and AUROC of ~56%. Accuracy in the model predicting prevailing cases at 9-10 yrs dropped to ~46% with a low AUROC value of ~45%.

Performance statistics of accuracy, precision, recall and the AUC are shown for the most accurate model obtained with deep learning optimized with Integrated Evolutionary Learning using a) multimodal features and b) only neural features. We used features obtained at 9–10 years of age to predict new onset cases of somatic symptom disorder at 11–12 years of age as well as all prevailing contemporaneous cases (9–10 yrs) and all prevailing cases at 11–12 years of age. Corresponding ROC curves may be viewed in Supplementary Figures 1 and 2.

In interpreting optimized multimodal predictive models for early adolescent SSD, we observed that new onset cases were predicted by whether the youth had ever received mental health or substance use (MH/SU) services, the parent's total burden of behavioral problems, total sleep disturbances, and whether the child had seen a clinician for a medical issue other than a regular checkup (Table 10a). While sets of specific predictors were not the same, overlap was observed among age groups. At 9–10 yrs, prevailing cases were predicted by parent withdrawal, inattention, somatic, and total behavioral problem scores, along with sleep disturbance, visits to a clinician for non-routine medical issues, and parent general behavior. For prevailing cases at 11–12 yrs, predictors included a relative's history of MH/SU services, parent aggressive and anxiety traits, sleep-wake transition disturbances, and total behavioral problems, as well as non-routine clinical visits and excessive somnolence disorders. Group-level importances for

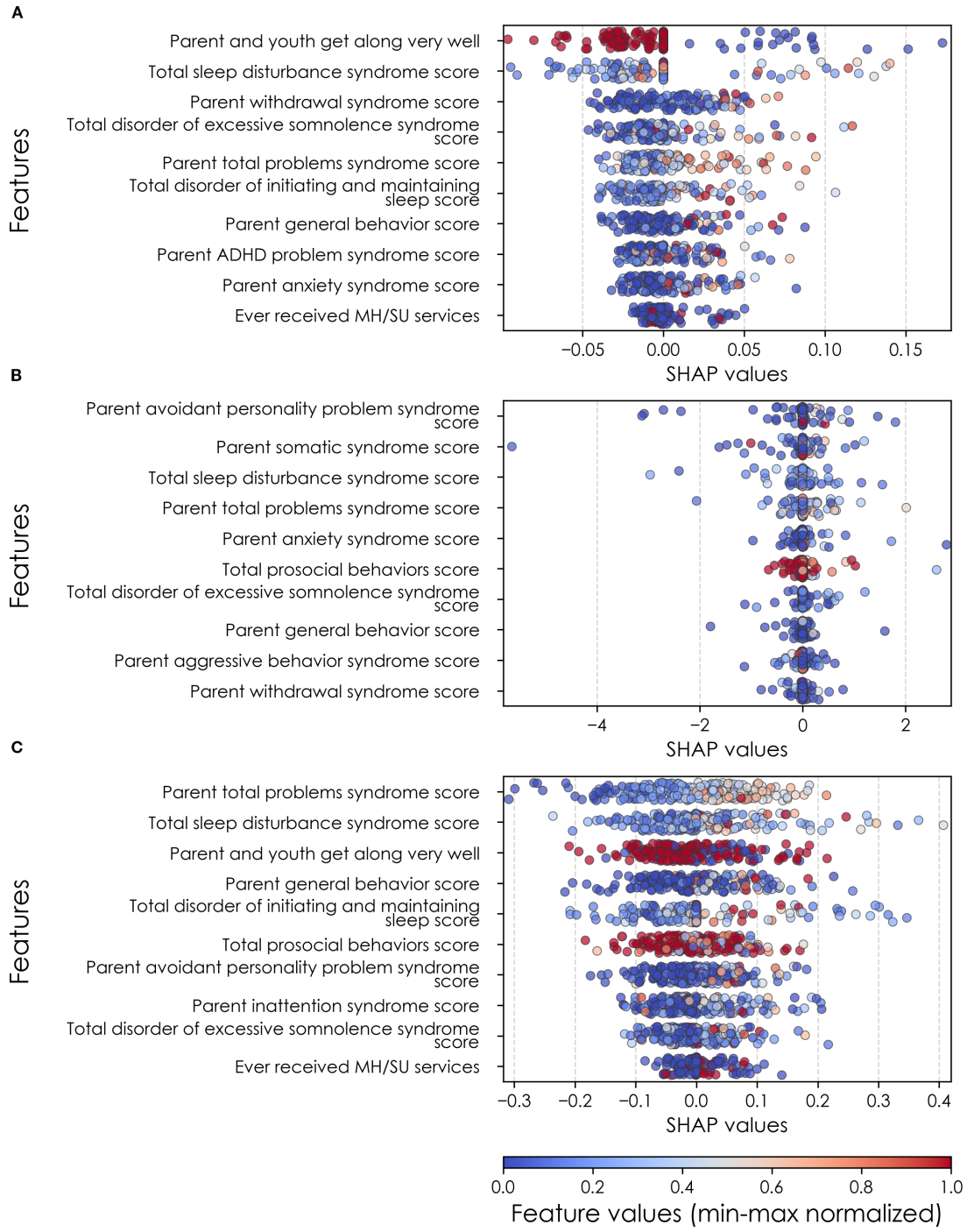

FIGURE 6
Individual-level importances of depression predictors in multimodal predictive models. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting depression **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient from low to high. MH, mental health; SU, substance use.

multimodal model predictors were in the range [0.05, 0.15] and the mean importance for each experiment in the range [0.07, 0.11].

Final predictors of cases of all prevailing cases of SSD at ages 9–10 and 11–12 years as well as new onset cases only at 11–12 years of age are shown for the most accurate models obtained using deep learning optimized with IEL obtained with a) multimodal features and b) only neural features. Final predictors are ranked in order of importance where the relative importance of each predictor is computed with the Shapley Additive Explanations technique and presented here averaged across all participants in the sample.

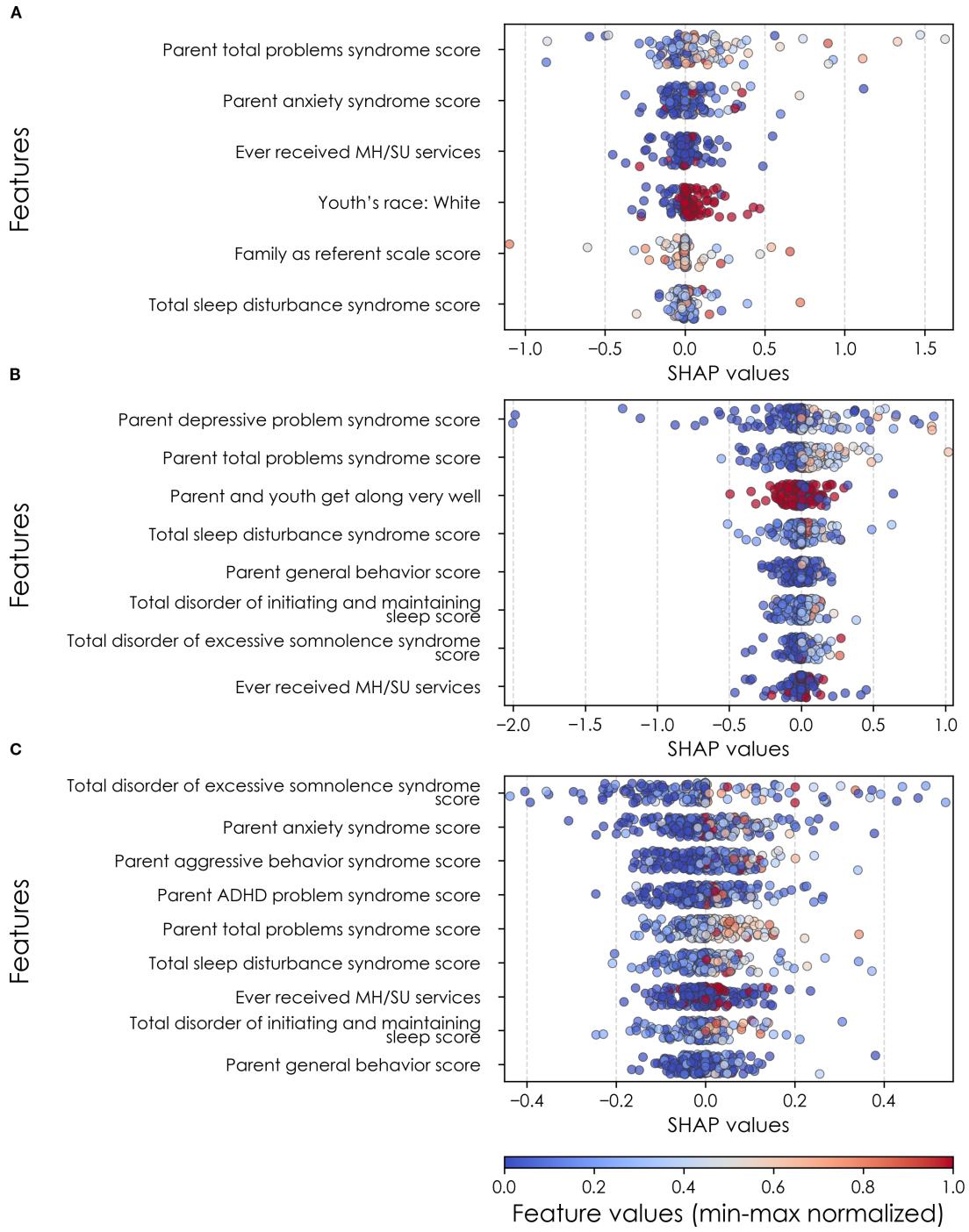

FIGURE 7
Individual-level importances of final predictors of anxiety in early adolescence. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting anxiety **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient. MH, mental health; SU, substance use; SST, Standard Stop Signal task; MID, Monetary Incentive Delay task; ROI, region of interest; FA, fractional anisotropy; LD, longitudinal diffusivity; WM, white matter; GM, gray matter.

Features in blue indicate an inverse relationship with depression verified with the Shapley method. MH = mental health; SU = substance use; SST = Standard Stop Signal task; MID = Monetary Incentive Delay task; ROI = region of interest; FA = fractional anisotropy; LD = longitudinal diffusivity; WM = white matter; GM = gray matter.

In neural-only models, we found that MID and nBack fMRI task features were emphasized in predicting new onset cases,

10.3389/fpsyt.2025.1487894

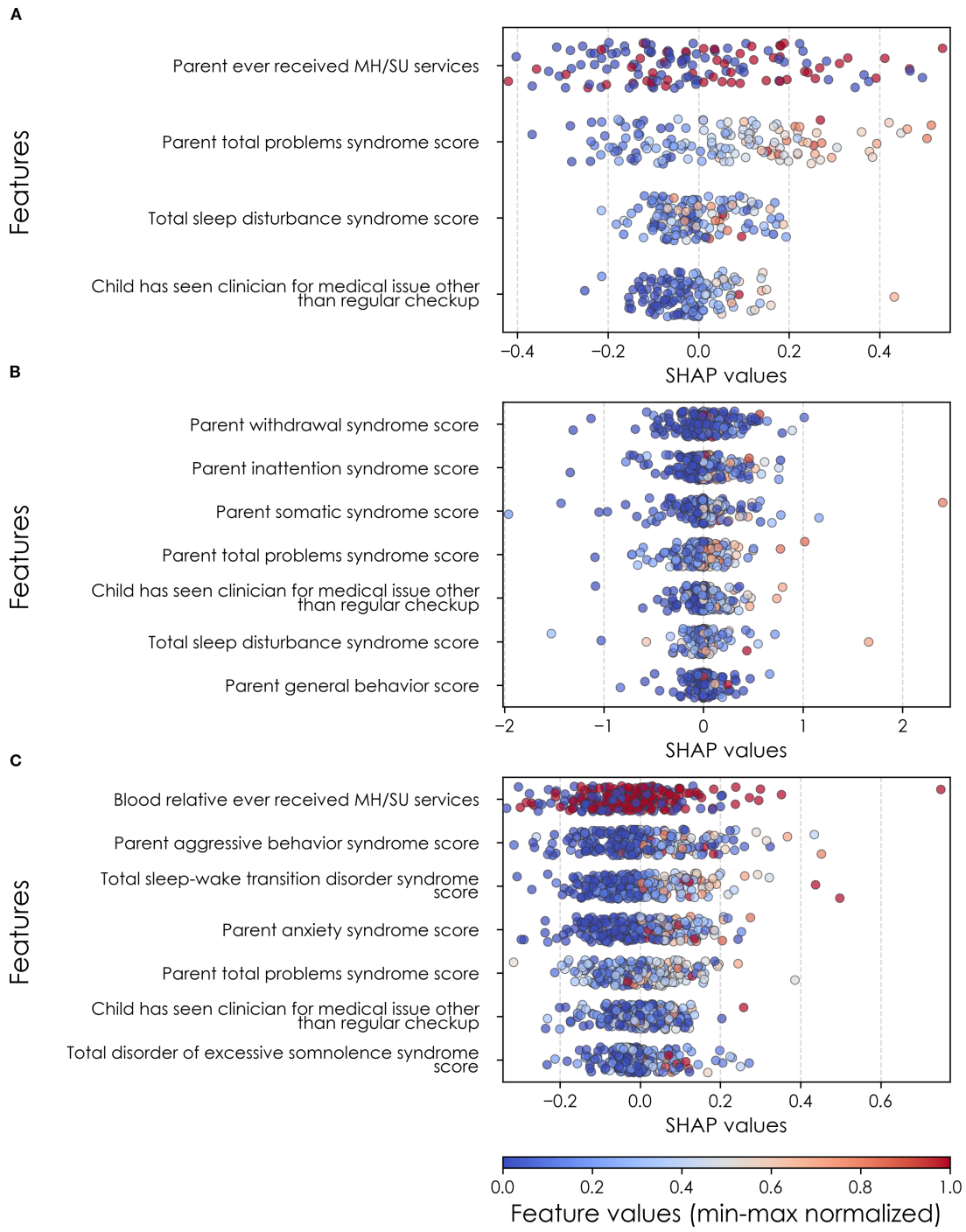

FIGURE 8
Individual-level importances of final predictors of somatic disorder in early adolescence. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting SSD **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient. MH, mental health; SU, substance use.

here emphasizing the frontal pole, left amygdala, cuneus and caudate. Important structural predictors of new onset cases were cortical volume in the right pars triangularis and contrast in the right lingual. As with new onset cases, final predictors of prevailing cases of SSD at 11–12 yrs centered on task fMRI metrics, again the MID with the addition of SST measures. Specific neural predictors of all prevailing cases at 11–12 yrs centered on the cuneus, insula and fusiform along with

10.3389/fpsyt.2025.1487894

TABLE 8 Final predictors of cases of depression in early adolescence.

a		
Age of case determination	Ranked final predictors	Importance
New onset at age 11–12 years	Parent and youth get along very well	0.028
	Total sleep disturbance syndrome score	0.027
	Parent withdrawal syndrome score	0.021
	Total disorder of excessive somnolence syndrome score	0.019
	Parent total problems syndrome score	0.017
	Total disorder of initiating and maintaining sleep score	0.017
	Parent general behavior score	0.017
	Parent ADHD problem syndrome score	0.015
	Parent anxiety syndrome score	0.014
	Ever received MH/SU services	0.008
	Mean	0.018
All cases at age 9–10 years	Parent avoidant personality problem syndrome score	0.13
	Parent somatic syndrome score	0.11
	Total sleep disturbance syndrome score	0.10
	Parent total problems syndrome score	0.08
	Parent anxiety syndrome score	0.06
	Total prosocial behaviors score	0.06
	Total disorder of excessive somnolence syndrome score	0.05
	Parent general behavior score	0.04
	Parent aggressive behavior syndrome score	0.04
	Parent withdrawal syndrome score	0.04
	Mean	0.07
All cases at age 11–12 years	Parent total problems syndrome score	0.08
	Total sleep disturbance syndrome score	0.06
	Parent and youth get along very well	0.05
	Parent general behavior score	0.05
	Total disorder of initiating and maintaining sleep score	0.05
	Total prosocial behaviors score	0.05
	Parent avoidant personality problem syndrome score	0.05
	Parent inattention syndrome score	0.04
	Total disorder of excessive somnolence syndrome score	0.04
	Ever received MH/SU services	0.02
	Mean	0.05
b		
Neural data type	Ranked final predictors	Importance
New onset at age 11–12 years	SST incorrect stop vs correct go contrast in left frontal pole ROI	0.30
	SST incorrect stop vs correct go contrast in left transverse temporal ROI	0.22
	T1 intensity in right ventricle ROI	0.21
	T1 white-gray contrast in right banks of the superior temporal sulcus ROI	0.16
	SST correct stop vs incorrect stop contrast in left frontal pole ROI	0.16
	SST any stop versus correct go contrast in left pars opercularis ROI	0.13
	FA in WM associated with cortical left anterior cingulate ROI	0.12
	T1 intensity in right accumbens ROI	0.11
	Mean	0.18
All cases at age 9–10 years	Correlation between sensorimotor mouth network and sensorimotor mouth network	0.18
	Correlation between cingulo-parietal network and sensorimotor hand network	0.17
	SST correct stop vs incorrect stop contrast in right pars triangularis ROI	0.16
	LD in GM associated with left banks of superior temporal sulcus ROI	0.14
	SST incorrect stop vs correct go contrast in right caudal middle frontal ROI	0.13
	LD in WM associated with right pericalcarine ROI	0.12
	LD in GM-WM contrast associated with right inferior temporal ROI	0.11
	Mean	0.14
All cases at age 11–12 years	FA in right lateral ventricle ROI	0.18
	MID anticipation of small loss vs neutral contrast in right inferior temporal ROI	0.16
	Correlation between auditory network and right accumbens area ROI	0.12
	Mean	0.15

(Continued)

TABLE 8 Continued

structural measures in parietal, putamen and orbitfrontal regions. In contrast, the final, optimized model predicting all prevailing cases at 9–10 yrs was dominated by connectivity metrics derived from rsfMRI, again emphasizing sensorimotor-control network connections (Figure 8).

When examined at the individual level, final predictors of SSD in each participant sample showed the same patterns as we observed in depression and anxiety. Individual-level predictor importances were widely dispersed, where typically the more important predictors exhibited wider dispersions (Figures 8, 11). Further, the dispersion of individual-level importances was greater in the more accurate multimodal models.

Similarly to depression and anxiety, individual-level importances of final neural predictors of somatic symptom disorder had a generally smaller variance in terms of feature values compared to the psychosocial predictors in the multimodal models, as indicated in Figures 8, 11. Their group-level importances were in the range [0.07, 0.20] where the disparity across important predictors were similar (Figure 8), suggesting no essential differences of explanatory power between them. The directionality of the relationship between predictors and SSD case status obtained

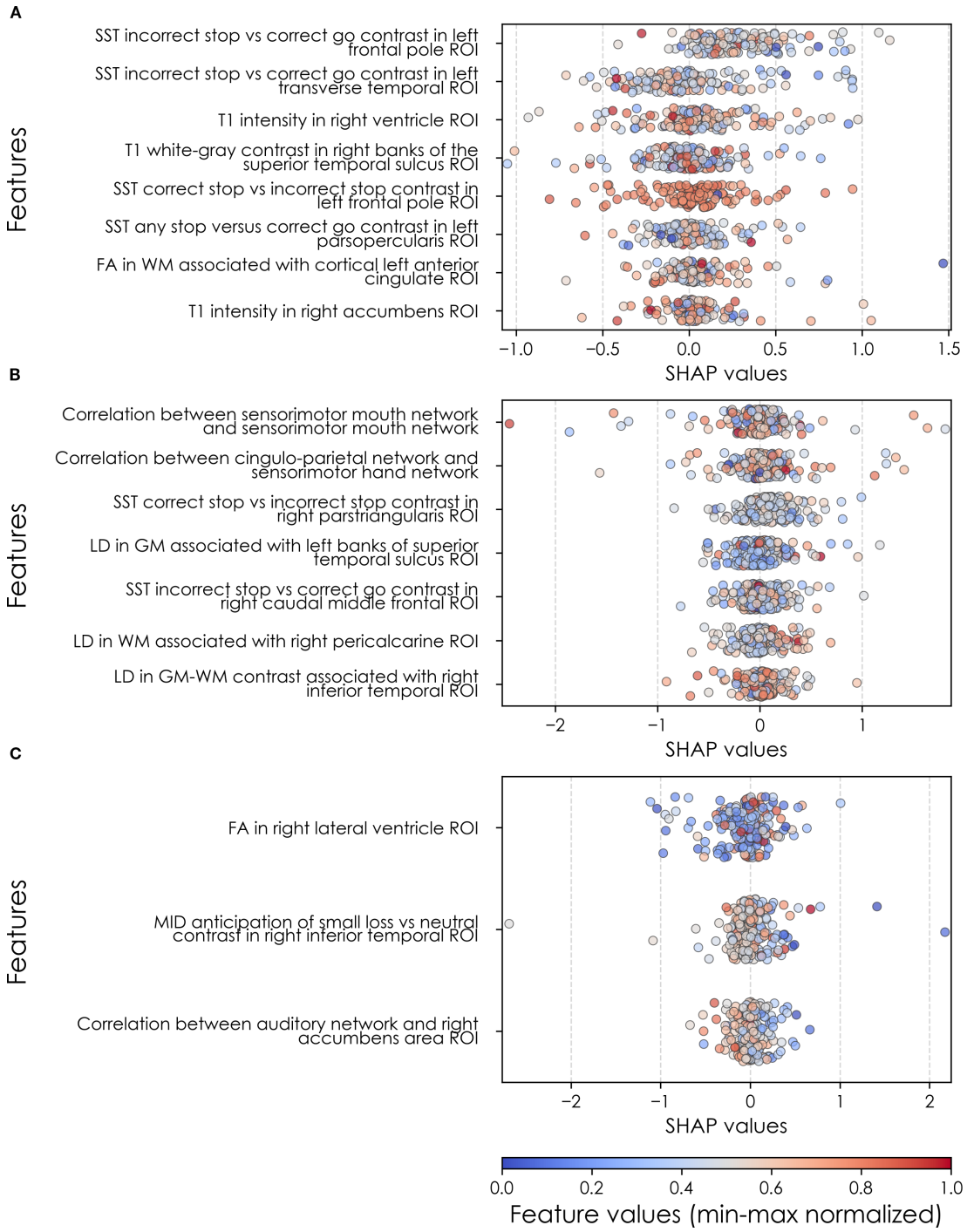

FIGURE 9
Individual-level importances of depression predictors in neural-only predictive models. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting depression **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient. MH, mental health; SU, substance use.

in these plots was further compared with coefficients obtained during LASSO regression for Coarse Feature Selection (Supplementary Table S3) and found to be in agreement. Further, Figure 5 visualizes the case study with two contrasting participants at the onset cases of 11–12 yrs, where parents' total problems syndrome score and their history of receiving mental health or substance use services are a strong risk factors for the children's SSD.

10.3389/fpsyt.2025.1487894

TABLE 9 Final predictors of cases of anxiety in early adolescence.

a		
Age of case determination	Ranked final predictors	Importance
New onset at age 11–12 years	Parent total problems syndrome score	0.18
	Parent anxiety syndrome score	0.10
	Ever received MH/SU services	0.07
	Youth's race: White	0.07
	Family as referent scale score	0.05
	Total sleep disturbance syndrome score	0.04
	Mean	0.09
All cases at age 9–10 years	Parent depressive problem syndrome score	0.15
	Parent total problems syndrome score	0.09
	Parent and youth get along very well	0.05
	Total sleep disturbance syndrome score	0.05
	Parent general behavior score	0.04
	Total disorder of initiating and maintaining sleep score	0.04
	Total disorder of excessive somnolence syndrome score	0.04
	Ever received MH/SU services	0.03
	Mean	0.06
All cases at age 11–12 years	Total disorder of excessive somnolence syndrome score	0.09
	Parent anxiety syndrome score	0.07
	Parent aggressive behavior syndrome score	0.06
	Parent ADHD problem syndrome score	0.06
	Parent total problems syndrome score	0.04
	Total sleep disturbance syndrome score	0.04
	Ever received MH/SU services	0.04
	Total disorder of initiating and maintaining sleep score	0.04
	Parent general behavior score	0.04
	Mean	0.05
b		
Neural data type	Ranked final predictors	Importance
New onset at age 11–12 years	Mean cortical sulcal depth in mm for left hemisphere	0.13
	Cortical thickness in mm of right pars triangularis ROI	0.12
	T1 intensity in brain stem ROI	0.12
	T1 white-gray contrast in left precuneus ROI	0.11
	MID anticipation of large vs small loss contrast in right parstriangularis ROI	0.11
	Mean	0.12
All cases at age 9–10 years	nBack negative face vs neutral face contrast in right postcentral	0.16
	SST correct stop vs incorrect stop contrast in 4th-ventricle ROI	0.14
	Correlation between cingulo-opercular network and left putamen ROI	0.13
	LD within right corticospinal	0.13
	Correlation between ventral attention network and left pallidum ROI	0.12
	SST correct stop vs correct go contrast in left inferior temporal	0.11
	Correlation between default network and sensorimotor hand network	0.10
	Correlation between cingulo-opercular network and ventral attention network	0.11
	Mean	0.10
All cases at age 11–12 years	Cortical thickness in mm of left transverse temporal ROI	0.15
	LD in WM associated with left cuneus ROI	0.14
	Correlation between none network and sensorimotor hand network	0.14
	Cortical thickness in mm of right pericalcarine ROI	0.13
	Correlation between cingulo-opercular network and dorsal attention network	0.13
	FA in WM associated with right entorhinal ROI	0.12
	LD in WM associated with left pericalcarine	0.12
	Correlation between fronto-parietal network and sensorimotor mouth network	0.11
	Cortical depth in left superior parietal	0.10
	Mean	0.13

(Continued)

Fairness subgroup performance analysis

To robustly evaluate the model performance from a fairness perspective, we conducted a subgroup performance analysis on the held-out test set, stratified by race and annual family income level. Performance statistics are reported in Supplementary Table S6.

For the stratification of race, 58.8-68.0% of the individuals in the held-out test set identified as White, 12.5-19.9% as Black, and 19.5-21.0% as other races. Accuracy scores for anxiety (79.3-84.0%) and somatic problems (83.3-86.2%) were largely consistent across different racial groups. However, in the case of depression, a larger disparity was observed: predictive accuracy was 92.6% for Black participants but 79.3% for participants categorized as other races, despite their similar sample sizes.

For the annual family income level stratification, 8.1-11.9% of the individuals in the held-out test set reported an annual income below $15,999, 9.0-17.1% earned $16,000-$34,999, 17.1-22.5% earned $35,000-$74,999, and 57.7-58.6% earned $75,000 or above. This stratification revealed larger performance disparities across income groups. The most extreme case was observed when predicting somatic problems, where the accuracy was 60.0% for the lowest income group compared to 91.3% for the $35,000-$74,999 group — a difference exceeding 30 percentage points.

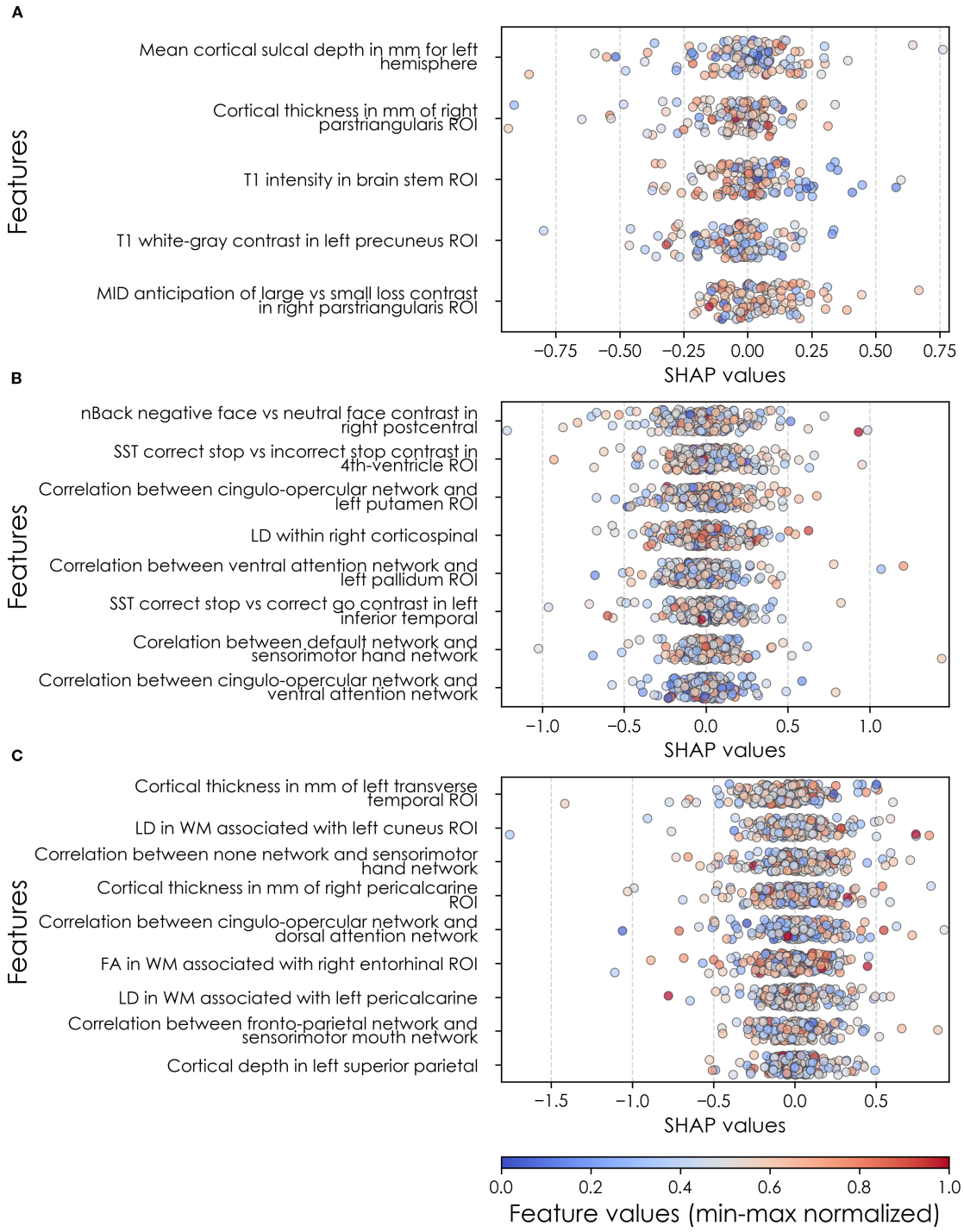

FIGURE 10
Individual-level importances of neural final predictors of anxiety in early adolescence. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting anxiety **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient. MH, mental health; SU, substance use; SST, Standard Stop Signal task; MID, Monetary Incentive Delay task; ROI, region of interest; FA, fractional anisotropy; LD, longitudinal diffusivity; WM, white matter; GM, gray matter.

TABLE 10 Final predictors of cases of somatic symptom disorder in early adolescence.

a		
Age of case determination	Ranked final predictors	Importance
New onset at age 11–12 years	Parent ever received MH/SU services	
	Parent total problems syndrome score	0.15
	Total sleep disturbance syndrome score	0.15
	Child has seen clinician for medical issue other than regular checkup	0.07
	Mean	0.06
		0.11
All cases at age 9–10 years	Parent withdrawal syndrome score	
	Parent inattention syndrome score	0.14
	Parent somatic syndrome score	0.14
	Parent total problems syndrome score	0.13
	Child has seen clinician for medical issue other than regular checkup	0.11
	Total sleep disturbance syndrome score	0.09
	Parent general behavior score	0.07
	Mean	0.06
		0.10
All cases at age 11–12 years	Blood relative ever received MH/SU services	
	Parent aggressive behavior syndrome score	0.09
	Total sleep-wake transition disorder syndrome score	0.09
	Parent anxiety syndrome score	0.08
	Parent total problems syndrome score	0.08
	Child has seen clinician for medical issue other than regular checkup	0.07
	Total disorder of excessive somnolence syndrome score	0.05
	Mean	0.05
		0.07
b		
Neural data type	Ranked final predictors	Importance
New onset at age 11–12 years	Cortical volume in mm^3 of right pars triangularis ROI	
	MID all anticipation of small reward vs neutral contrast in 4th-ventricle ROI	0.20
	nBack positive face vs neutral face contrast in right cuneus ROI	0.17
	T1 white-gray contrast in right lingual ROI	0.17
	MID all anticipation of small reward vs neutral contrast in right frontalpole ROI	0.14
	MID all loss positive vs negative feedback contrast in left amygdala ROI	0.14
	nBack negative face vs neutral face contrast in right caudate ROI	0.14
	Mean	0.14
		0.16
All cases at age 9–10 years	Correlation between sensorimotor mouth network and sensorimotor mouth network	0.16
	Correlation between dorsal attention network and sensorimotor hand network	0.15
	Cortical sulcal depth in mm of right postcentral ROI	0.15
	Average correlation between auditory network and cingulo-opercular network	0.11
	FA in GM associated with left inferior temporal ROI	0.11
	T1 within corpus callosum mid-anterior	0.11
	Correlation between default network and retrosplenial temporal network	0.09
	T1 in WM in left transverse temporal	0.13
	Correlation between retrosplenial temporal network and sensorimotor mouth network	
	TD in GM associated with left entorhinal ROI	
	Mean	
All cases at age 11–12 years	SST incorrect go vs correct go contrast in left lateral-ventricle ROI	
	Cortical area in mm^2 of left inferior parietal ROI	0.12
	Diffusivity within left putamen ROI	0.12
	MID all anticipation of large vs small loss contrast in left cuneus ROI	0.12
	SST incorrect go versus incorrect stop contrast in left cerebellum cortex ROI	0.11
	MID all anticipation of small reward vs neutral contrast in right insula ROI	0.11
	MID all reward positive vs negative feedback contrast in right lateral ventricle ROI	0.10
	MID all loss positive vs negative feedback contrast in right fusiform ROI	0.09
	Depth of right orbitofrontal	0.08
	Cortical thickness in mm of right lingual	0.07
	Mean	0.07
		0.10

(Continued)

TABLE 10 Continued

Baseline comparison

We compared our IEL approach with three traditional ML modeling methods: logistic regression, random forest, and SVM. The performance statistics, defined in terms of accuracy, precision, recall, and AUROC, of these models, trained with the multimodal feature sets, are reported in Supplementary Table S7. The results demonstrate that the baseline models achieved highly comparable performance with IEL consistently. The average accuracy (across targets and cases of determination) of IEL is 86.1% whereas the baseline models, which include roughly 10 times more features, have an average accuracy of 87.5-88.7%. Student t-tests verified that IEL does not have a statistically significantly different accuracy

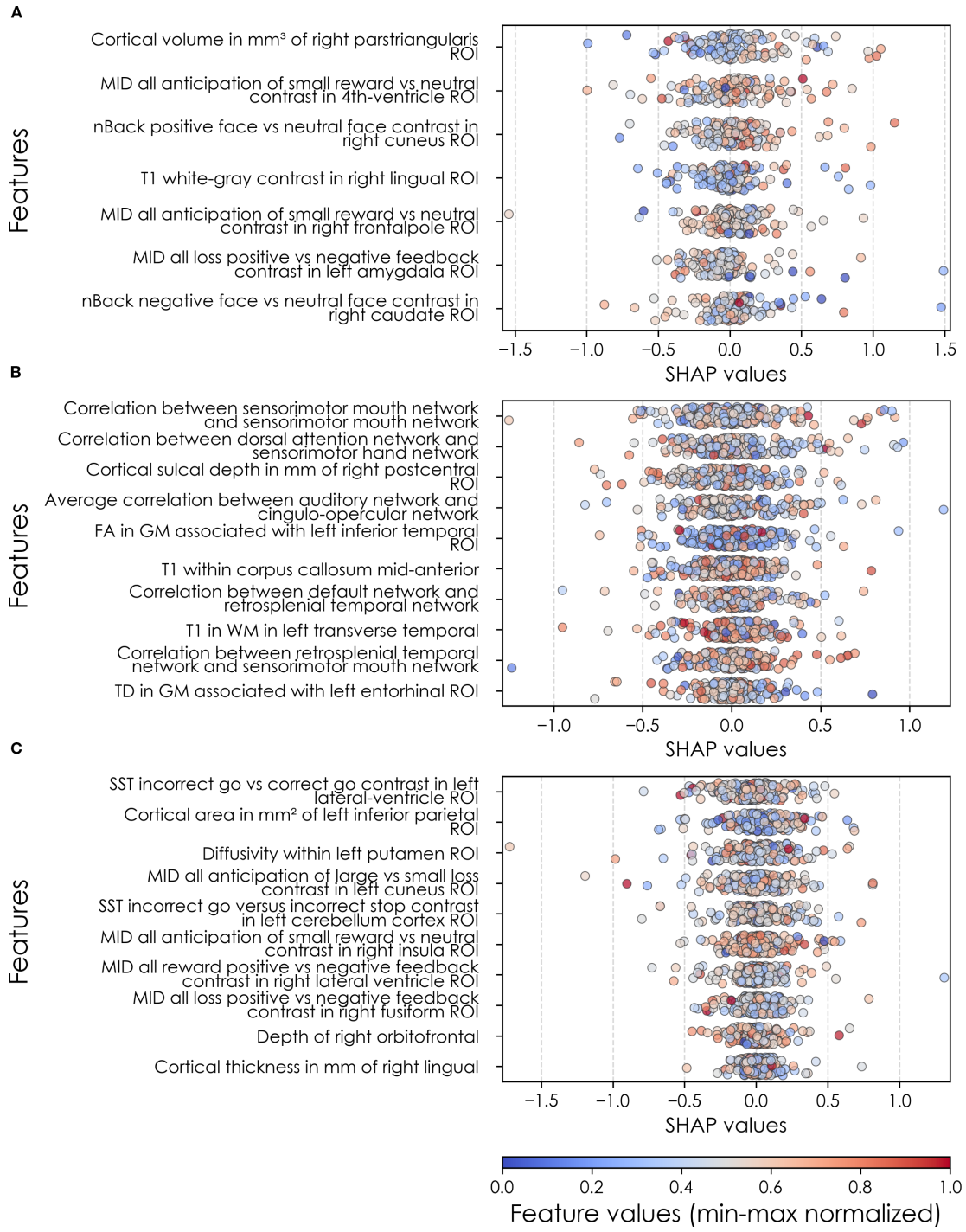

FIGURE 11
Individual-level importances of neural final predictors of somatic symptom disorder in early adolescence. Summary plots are presented of the importance of each final predictor (computed with the Shapley Additive Explanations technique) on an individual subject level to predicting SSD **(A)** with new onset at 11–12 yrs; **(B)** in all cases at 9–10 yrs; and **(C)** in all cases at 11–12 yrs. The color gradient represents the original value of each feature (metric) where red = high and blue = low. Discrete (binary) features appear as red or blue, while continuous features appear as a color gradient. MH, mental health; SU, substance use.

performance compared to logistic regression (p-value = 0.36), random forest (p-value = 0.54), or SVM (p-value = 0.23).

Discussion

Common and specific themes across internalizing disorders

We analyzed ~6,000 candidate predictors from multiple knowledge domains (cognitive, psychosocial, neural, biological) contributed by children of late elementary school age (9–10 yrs) and their parents and constructed robust, individual-level models predicting the later (11–12 yrs) onset of depression, anxiety and SSD. Leveraging an optimization pipeline that included AI-guided automated feature selection allowed us to extend prior work by analyzing a wider variety of predictor types and ~40x more candidate predictors than previous comparable ML studies. A common pre-processing and analytic design across all three internalizing disorders in the same youth cohort allows the direct comparison of results to elicit their diagnostic specificity and identify common themes. In addition, we wanted to quantify the relative predictive performance of multimodal vs neural features and examine the relationship between predictor importance and model accuracy. To our knowledge, this is the first ML study in adolescent internalizing disorders to include multiple types of neural predictors (rsfMRI connectivity; task fMRI effects; diffusion and structural metrics), analyze >200 multimodal features and quantify the relationship between predictor importance and accuracy. The iterative feature sampling approach adopted in our genetic algorithm offers additional robustness to this quantification by reducing the risk that predictive accuracy hinges on a single subset or on correlated predictors retained in the pre-processing phase.

Comparing across results, we found that the relative predictive performance of our models varied according to the specific disorder and type of predictor (psychosocial vs neural). Deep learning optimized with IEL rendered robust individual-level predictions of all three internalizing disorders with AUROCs of 86-99% and 79-94% accuracy. Precision and recall were also consistently ≥~80% with scattered exceptions in precision (new onset anxiety: 76%) and recall (new onset depression: 78%; prevailing SSD at 11–12 yrs: 78%). Our primary focus was in predicting future, new onset cases of each internalizing disorder in early adolescence. We found that new onset cases of depression could be most reliably predicted (AUROC ~93%), followed by SSD (AUROC ~90%) and anxiety (AUROC 86%).

An important result is that we found that predicting early adolescent internalizing disorders with multimodal features resulted in substantially better performance than exclusively neural-based models, and that psychosocial predictors were preferentially selected in multimodal modeling. Our pipeline includes automated feature selection with a genetic algorithm (IEL) that progressively selects among features as it learns how to optimize predictive models over a principled training process (typically ~40,000 models). Cognitive, neural and biological features failed to outcompete psychosocial features in training with multimodal. Further targeted experiments specifically assessed the standalone predictive ability of multiple neural feature types derived from MRI. These experiments demonstrated that neural-only models achieved a close to 50% performance, sacrificing 24-45% performance compared to the multimodal models, across statistics (accuracy, AUROC, precision, recall). While little extant research has directly compared psychosocial to neural features in youth internalizing disorders, our results are congruent with studies that have used multimodal feature types including MRI metrics (18, 19). Our design extended prior work by allowing us to examine more and wider feature types and disorders and the prediction of new onset vs prevailing cases. Neural-only models of new onset cases achieved superior performance to other participant samples and selectively comprised task fMRI and structural metrics, though more neural feature types (rsfMRI connectivity, diffusion-based) were available for selection, suggesting structural and task fMRI neural features may have particular promise in predicting adolescent onset of internalizing disorders.

Specific sets of final predictors for each disorder and participant sample were unique and differentiated both a) depression, anxiety and SSD from each other and b) future new onset from all prevailing cases. However, parental levels of various types of problem behaviors and youth sleep disturbances appeared as cross-cutting, higher-level themes. All three internalizing disorders showed commonality in parent-related psychopathology measures, with anxiety- and attentional-related difficulties assorting as predictors across different participant samples. Notable disorder-specific predictors included parent level of somaticizing to their child's SSD. Taken together, our results demonstrate that parent problem behavioral traits are important drivers of internalizing disorders in early adolescence and that the specific parental traits observed when their child is 9–10 yrs may be useful in discriminating whether their child will go on to develop depression, anxiety or SSD. This phenomenon suggests intergenerational transmission, though our design cannot determine whether this is underpinned by inheritance, parent-youth styles of relating or other factors, though the presence of externalizing parental behaviors in predicting the later onset of an internalizing disorder in the child suggests that more than inheritance is at work. Here, our results congruent with the small number of comparable ML studies that have included parental traits as candidate predictors, where parent total behavioral problems and poor maternal relationships were leading predictors of depression (15, 61). These findings are also consistent with developmental frameworks, such as Bronfenbrenner's bioecological model, which emphasizes the role of proximal family factors in shaping child outcomes (26), and Ciccheti and Rogosch's developmental psychopathology framework, which underscores how multiple risk pathways can converge on internalizing outcomes (27). Similarly, Sameroff's transactional model highlights the reciprocal inference of parents and children over time, further supporting the central role of parental psychopathology in developmental trajectories (28).

Next, sleep disturbances may affect up to ~40% of elementary school age children and youth with both internalizing and externalizing disorders are at elevated risk (62, 63). We found

that sleep disturbances in the late elementary school age group (9-10yrs) predicted the later (11–12 yrs) onset (anxiety, SSD) and prevalence (depression) of internalizing disorders, congruent with recent research showing that disturbed or short duration sleep predicts later internalizing symptoms (64–67). Here, our findings add to a growing body of work suggesting sleep disturbances may be important intervention targets in elementary school age youth to reduce the later burden of internalizing symptoms (67). Beyond empirical associations, sleep disturbance has also been conceptualized as a transdiagnostic vulnerability process in developmental psychopathology frameworks. Disruptions in sleep and arousal regulation may impair affect regulation, cognitive control, and stress reactivity, thereby increasing susceptibility to internalizing disorders (29–31).

Recent research in association-based studies has suggested that effect sizes in neuroimaging studies of psychopathology and cognitive traits are often inflated, particularly in smaller participant samples, resulting in generalization failure (68). Our prior work in predicting externalizing disorders has similarly shown that neural predictors tend to underperform (69). Accordingly, we investigated predictor importance at both the group and individual level and its relationship with model performance in generalization testing, observing a strong relationship between predictor importance and accuracy across experiments. In individual experiments, psychosocial predictors in multimodal models exhibited generally greater variances in feature values than those in neural-only experiments, even after extensive optimization and principled feature selection. Collectively, these results suggest that the more restricted variability of neural predictors among individuals were at least related to their weaker performance in predicting cases using artificial neural networks. Future work will be required to determine whether these phenomena are seen in other disorders and participant samples (particularly other developmental periods) and if other types of neural features (for example, connectivity features obtained from data-driven rather than ROI methods) could fare better in predicting cases of internalizing disorders.

Depression

Depression is a common and growing problem in adolescence which elevates later risk for suicide, poor educational outcomes and substance use (70). In the present study, we focus on early onset cases of depression i.e. those which onset or are present at 11–12 yrs. Most prior work in early onset depression has examined psychosocial predictors at the group level, linking it to sleep disturbances, childhood adverse events (neglect, abuse, loss of parent), familial depression and pubertal changes (71–77) Longitudinal neuroimaging studies of the onset or course of depression in adolescence are relatively plentiful and have ranged across a variety of MRI modalities (78). Similarly, these have typically been group-level studies employing traditional multivariate predictive methods in a single neuroimaging modality and small number of ROIs, sometimes in small samples. Results have been inconsistent. In structural MRI, subcortical regions (especially hippocampal) have been most intensively studied with mostly negative results, though there is some evidence for smaller accumbens and insula volume and equivocal results for OFC regions (79–84). In fMRI, reward and emotion processing have been most intensively studied. A number of studies have demonstrated differential reward-related activity in the ventral striatum (85–89), though these studies are nearly all from later adolescence. In early adolescence, Morgan et al. found the inverse was the case (90). In emotion processing, increases or decreases in ACC activity have predicted adolescent depression onset (91–93).

More recently, a number of ML studies have performed prospective prediction of adolescent depression incorporating larger numbers of candidate predictors, either psychosocial and/or neuroimaging. To our knowledge, our study represents only the second time multimodal (including neuroimaging) candidate predictors have been analyzed at the individual level using ML to prospectively predict depression onset in adolescents, and the first time in early adolescence. With an AUC of ~0.90, we achieved performance comparable with a single prior deep learning study and superior to that obtained using logistic regression or support vector machines (SVM) (18, 61, 84–86). We are not aware of other prior ML studies that have directly compared the ability of multimodal vs neuroimaging predictors in adolescent depression or incorporated more than one type of neuroimaging metric.

Our AI-guided optimization pipeline preferentially selected psychosocial features to predict early onset adolescent depression after analyzing thousands of multimodal candidate predictors. Multimodal models achieved 27-39% better performance over all metrics than neural-only models. However, at ~0.57 AUROC, our neural-only deep learning model achieved performance inferior to multimodal models in other studies using different ML methods (logistic regression, SVM). Several recent large-scale ML prospective predictive studies of youth depression have examined the predictive performance of nonlinear combinations of candidate predictors at the individual level. In youth aged 15 yrs, Rocha et al. trained penalized logistic regression models with 11 psychosocial metrics finding that school failure, social isolation, involvement in physical fights, drug use, running away from home, and maltreatment predicted depression onset at 18 yrs, achieving AUROC 0.79 in the baseline dataset and 0.59 and 0.63 in external validation datasets (94). Foland-Ross et al. used cortical thickness metrics to predict new onset adolescent depression with 70% accuracy, with thickness of the right precentral and medial OFC and left ACC and insula representing the most important features (84). Most recently, two important large scale ML studies utilized multimodal candidate predictor sets. Toenders et al. applied penalized logistic regression to 69 phenotypic and 76 structural MRI metrics in youth aged 14 yrs from the IMAGEN dataset, testing for generalization in a held-out set to achieve 0.72 AUROC and 66% accuracy (78). Depressive symptoms at baseline, neuroticism, cognition, supramarginal gyrus surface area, and stressful life events were most predictive of later new onset depression. Xiang et al. surveyed 188 psychosocial and rsfMRI connectivity candidate predictors collected at 9–10 yrs and

empirically selected based on prior literature to predict depression trajectories (computed with latent class analysis) through 11–12 yrs in the ABCD cohort, with deep learning achieving best performance (61). This study is perhaps the most comparable to our own methodologically and achieved similar AUROC (~0.90) and accuracy (87% vs ~82, ~86%), though precision (0.45) and recall (0.44) were lower. Total sleep disturbance, parent total behavioral problems, financial adversity, ventral attention-left caudate and dorsal attention-left putamen connectivity and school disengagement were the most important predictors of depression trajectories. Thus, we obtained thematically concordant results with prior research in identifying parental problem behaviors of various types and sleep disturbances being important predictors of early adolescent depression. However, our work differs in not identifying other types of childhood adverse experiences, cognitive traits and pubertal status as being as important to final, optimized models. In new onset depression, we found that the most important predictors were the tenor of the parent-child relationship, parent withdrawn and inattentive traits and sleep disturbances. In contrast, parent avoidant and more general metrics of behavioral issues specifically drove the prediction of all prevailing cases at 11–12 yrs.

We believe that this is the first time that multiple neuroimaging feature types have been used to predict new onset depression in adolescence in a neural-only model. Thus, it is particularly intriguing to note that the onset of early adolescent depression was predicted by multiple task fMRI effects – but that these centered on the SST (which measures response inhibition) rather than the MID (reward processing). We found rather that MID effects were emphasized in predicting anxiety and in particular SSD – and it has been previously noted that almost no longitudinal fMRI studies in adolescent depression directly compare anxiety and depression in the same sample (78). In our neural-only models, results are concordant with existing literature in highlighting fronto-temporal ROIs but our algorithms preferentially selected effects from the SST over the MID. The SST is a test of inhibition of prepotent responses and has been extensively studied in externalizing disorders (where there is a positive relationship) but less in the internalizing disorders. However, ex-scanner studies in children with internalizing behaviors and adults with depression using the SST show longer reaction time in patients with recent work associating response inhibition deficits in children with rumination traits (95–97). Future work may consider exploring SST task-related effects in response inhibition further in adolescent depression.

Anxiety

Anxiety is among the most common mental health disorders affecting adolescents and adults. Among the internalizing disorders, it is the condition most clearly centered on early adolescence, with a median age of onset of 11 yrs. Many psychosocial, demographic and cognitive risk factors have been associated with the development of clinical anxiety including early life temperamental traits such as anxiety sensitivity, neuroticism and anxious temperament. Thus, the formulation of prospective predictive models that can discriminate among these factors and provide reliable, individual-level predictions of anxiety onset in early adolescence is of particular relevance. However, few ML studies have predicted future anxiety in adolescence. To our knowledge, this is the first ML study to predict future anxiety in early adolescence and the first to use multiple neural features types. In important prior multimodal work, Chavanne et al. compared the ability of psychosocial vs neural features to predicting anxiety cases at 18–23 yrs in the IMAGEN cohort with 14 gray matter volumetric measures and 13 clinical metrics measured at 14 yrs using a majority voting algorithm comprising Logistic Regression, SVM and Random Forest classifiers. In the multimodal model, an AUROC of 0.68 was obtained with neuroticism, hopelessness, emotional symptoms and family factors contributing most to the prediction and volumetric differences in the periaqueductal gray, amygdala, ACC and subcortical regions making lesser contributions. With neural features alone, AUROC dropped to 0.52 whereas with psychosocial features alone it improved to 0.69.

Here, we demonstrate that new cases of anxiety at 11–12 yrs can be reliably (AUROC ~86%; accuracy ~79% and precision ~84%) predicted with deep learning optimized with IEL and that these predictive models differ from depression and SSD. As in the developmentally older IMAGEN cohort, our analysis in the younger ABCD cohort found that multimodal features predict the onset of anxiety better than neural-only features with a substantial differential of 24-40% across performance statistics. We found that new onset cases of anxiety in early adolescence were predicted by elevated parental mental health issues, sleep disturbances and the child having come to prior clinical attention. It is noteworthy that elevated parental anxiety trait scores was a specific predictor of their child's anxiety. While the child having White race appeared as an important predictor, we emphasize that predictive models are not mechanistic and this factor could easily represent diagnostic frequency. Our results are congruent with the literature and suggest that elevated parental anxiety and the total burden of parental behavioral issues and child sleep disturbances interact in a nonlinear manner to predict the onset of later clinical anxiety in early adolescence. While there was thematic overlap among our different anxiety models (parent problem behaviors, sleep disturbances) this particular set of factors was specific to new onset cases. While parent depressive behaviors were a final predictor of contemporaneous cases at 9–10 yrs, they did not predict new onset cases at 11–12 yrs.

Somatic symptom disorder

Somatic behavioral problems refer to the presence of one or more physical symptoms accompanied by excessive investment (time, emotion, behaviors) in the symptom(s) that results in significant distress or dysfunction. The diagnosis of SSD emphasizes symptom-based impairment in daily life. Peri-adolescence is an important period when SSD onsets and rises towards higher adult rates. Prior research, including prospective

studies, has frequently implicated family functioning including parents' own levels of physical and mental health complaints and parent somatic problems as well as parental divorce, illness or death, childhood traumatic experiences and insecure attachment (98–103). Work examining adolescent predictors of subsequent trajectories of somatic symptoms have identified the quality of parent-youth relationships, parenting stress and youth bullying, school dissatisfaction and lower intelligence level symptoms as important predictors (104–108). The genetic component appears to be small, albeit studies are limited (109). Research focused on the cognitive-affective neural basis of somatic problems using task fMRI has linked group-level differences in para/hippocampal, ACC, insula, brainstem and lateral prefrontal regions to effects in negative expectancy, attentional bias and pain catastrophizing (110–116). Fewer neuroimaging studies have investigated circuit abnormalities in somatic problems, though rsfMRI studies have implicated increased brainstem, caudate, thalamus and ACC activity and decreased lateral prefrontal activity in adults (117, 118). In a cross-sectional study in the ABCD cohort, Dhamala et al. found disrupted temporo-parietal, default mode, dorsal attention and control-limbic functional connections using rsfMRI data from 9–10 yrs to predict CBCL somatic problem scores at the same age (119).

Our findings contribute to this growing body of work in several ways. Firstly, prospective predictive studies of somatic problems have typically focused on either psychosocial (particularly family- or adversity-related measures) or neural predictors. In the present study we analyzed nearly 6,000 multimodal predictors of many types (including cognitive and non-neural biological metrics), allowing us to assess their relative predictive ability holistically. In these multimodal models, we found that psychosocial predictors were preferred over neural, cognitive and biological metrics. Secondly, the richness of parent and family-related metrics in the ABCD sample allowed us to consider a larger range of psychosocial predictors than has typically been available to earlier studies of somatic problem symptoms in youth. We found that parent level of somatic problem behaviors (9–10 yr prevailing cases) and internalizing as well as externalizing traits were preferentially selected as predictors over other family-, school- or peer-related candidate predictors such as bullying, parent stress or early adverse experiences. In all participant samples, parent somatic or internalizing problem behaviors interacted with sleep disturbances. Of note, whether a specific predictor of somatic problems in new onset cases at 9–10 yrs and cases at 11–12 years was whether the child was seen for a medical issue other than a regular checkup. These findings comport with earlier work and further suggest that childhood patterns of clinical use and sleep disturbances and elevated levels of parent somatic traits may be helpful in assessing youth risk for somatic problem behaviors. Similarly, the wide range of neuroimaging measures available allowed us to assess nearly 5,000 different neuroimaging metrics over multiple modalities to predict somatic problem behaviors in youth. While these models were not as robust as multimodal models (AUROC ~0.45-0.56), they are congruent with extant research in centering on temporal, frontal and cingulate regions and attentional network connectivity. Our work additionally highlighted the insula, a region long known to be involved in interoception and pain processing. Interestingly, effects in these regions during the MID task involving reward processing and loss anticipation were emphasized in predicting new onset cases of somatic problems in contrast to anxiety, where they centered on loss anticipation only. While we are not aware of prior work using the MID task in somatic problem behaviors, this may be an interesting line of future inquiry given a cardinal feature of somatization is the amount of valence and/or investment given to physical symptoms. Overall, we found that structural, task and rsfMRI were useful modalities in predicting somatic problems in early adolescence but diffusion imaging made less of a contribution.

Fairness subgroup performance analysis

The ABCD study was designed to approximate the demographic characteristics of U.S. children through stratified probability sampling of schools across 21 U.S. sites. Although it does not guarantee full representativeness (120), some prior studies have argued for its national representativeness (121, 122). In light of the performance statistics reported in Supplementary Table S6, there exist visible disparities across subpopulations, particularly when stratified by annual family income levels. These findings highlight the need for future research to investigate the underlying sources of these disparities and to explore fairness-aware approaches that can promote more equitable predictive performance. While such fairness optimization is beyond the scope of the present study, our results provide a foundation for subsequent work to address these challenges.

Baseline modeling analysis

While the IEL technique performs only comparably to the traditional ML modeling techniques, it has clear deployment advantages — yielding more interpretable results due to its use of fewer features, an enhanced stability through the optimization process, and enabling individual-level explanation due to its incorporation of SHAP values. IEL embeds a highly explainable solution to the traditional ML modeling problems, offering the operator the ability to visualize increasing model efficiency over iterations to identify the optimal solution, where this solution is parsimonious. Ultimately, parsimonious models are better suited for clinical risk stratification purposes since they require the *de novo* collection of less data in the field.

By contrast, deep-learning models and genetic algorithms for hyperparameter optimization generally require greater computational resources for training. That said, prediction is relatively efficient once the model is trained. For example, prediction with our deep-learning models typically completes in less than a second, owing to the computation efficiency of the well-optimized PyTorch package. This tradeoff between training time and model interpretability is reasonable in clinical practice, where

models are trained infrequently but deployed repeatedly to provide rapid, interpretable predictions in single shot learning for individual patients at the point of care. Thus, IEL framework balances practical feasibility with the clinical need for transparent, individualized risk assessments.

Predictive models and their future value for risk stratification

Individual-level predictive models such as we present in this paper can be valuable in clinical practice for their role in providing the core of risk stratification algorithms, which calculate the amount of risk an individual has for a specific condition. Risk stratification is a multi-stage developmental process where the first step is building predictive models with robust positive class discrimination. This is typically followed by deciding on an intervention, coupling this intervention with the predictive model to form a decision support tool and testing this tool in the clinical population. Using a robust risk prediction model, clinicians can stratify an individual's relative risk level to initiate preventive monitoring or supportive interventions at an earlier stage. While developing and validating a fully deployable decision support tool for risk assessment is beyond the score of the current study, our model establishes a foundation for future work aimed at integrating predictive risk score into clinical workflows. Our findings also have potential implications for the development of risk stratification tools in child and adolescent mental health. Predictive models such as IEL could be integrated into clinical or educational settings to classify youth into relative risk tiers (e.g., high, moderate, or low risk) for later internalizing disorders. Such stratification could enable more efficient allocation of limited resources, with higher-risk individuals receiving targeted screening, preventive support, or referral for early intervention.

An important consideration in translating predictive models into practice is the precision-recall trade-off. While maximizing sensitivity is valuable for identifying youth at risk of internalizing disorders, this inevitably comes at the cost of reduced precision, leading to false positives. In clinical contexts, false positives may carry consequences such as unnecessary monitoring, referrals, or anxiety for families. These potential drawbacks must be weighed against the benefits of early identification, particularly when interventions are low-risk or preventive in nature. Accordingly, predictive models such as ours should be regarded as adjunctive decision-support tools that complement, rather than replace, clinical expertise in assessing risk.

Validation in external, independent samples will be required in future work to strengthen the current analysis. Currently, to our knowledge, there is no publicly available dataset with comparable combination of sample size, longitudinal depth, and breadth of measurement as the ABCD Study, making external validation challenging within the scope of this work. Also, although subsequent ABCD releases (e.g., Release 6.0) provide additional data points, they involve the same cohort of participants at a later age point while the present analysis is specifically focused on early adolescence. As the ABCD Study continues to collect and release longitudinal data, future research will benefit from extending predictive analyses into mid- and late-adolescence. Such work will allow evaluation of whether early predictors identified in this work remain stable across development or whether new risk factors emerge during later stages of adolescence. These additional analyses will be critical for understanding the developmental timing of risk pathways and for refining prediction models across the adolescent period.

Additionally, expanding the feature space to include other modalities can potentially enhance the predictive power. For instance, sensor data recorded by electronic wearables or mobile phones can capture how the children's daily activities, screentime usage, and exercise levels impact the risk of internalizing disorders. Another route to improve the practicality of our adopted algorithm is to explore other hyperparameter optimization strategies. Similar to genetic algorithms which have a strong theoretical underpinning, Bayesian optimization techniques and bandit-based methods (123) are also feasible alternatives to streamline to model training process.

Limitations

This study uses secondary data from the ABCD study and we were therefore unable to control for any bias during data collection. While the ABCD study strived for population representation, there is a mild bias toward higher-income participant families of white race in the early adolescent cohort. Thus, the ABCD study may not fully represent all racial, ethnical, and socioeconomic groups within the U.S. population, or even to the broader non-U.S. populations. While our findings in the fairness subgroup analysis indicate disparities in predictive performance, investigation of their causes and mitigation strategies is beyond the present study but highlights important directions for other researchers to explore these important questions in the future research. On the other hand, data is not available prior to baseline (age 9–10 years) assessment and we cannot conclusively rule out that youth participants met criteria for depression, anxiety or somatic problems prior to this age but not at baseline assessment at 9–10 years of age. Thus, it is possible that certain cases coded as 'new onset' at 11–12 years of age in our analysis could have met clinical criteria ≤8 yrs but were in remission at 9–10 yrs. In the present study, we defined cases as any individual meeting ASEBA clinical thresholds in the CBCL subscale scores of interest and did not exclude participants who thereby met criteria for other conditions. Thus, co-morbidity may be present in the experimental samples as is common in clinical populations and in most research studies in early adolescence. While we used nearly 6,000 variables available in the ABCD dataset, our study is not exhaustive. It is possible that different results could have been obtained if more or different candidate predictors were included. For example, rsfMRI data includes metrics from ROI-based parcellations but not a data-driven method such as ICA. We focused on rigorous internal validation strategies by including strict separation of training and test sets, and evaluation on the holdout test set that was never used in model training or validation, a gold standard in ML that ensures an unbiased evaluation of generalizability. However, prospective external validation using a

dataset other than ABCD can further improve generalizability of our analysis.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data is available from the NBDC Data Hub Requests to access these datasets should be directed to https://www.nbdc-datahub.org/.

Ethics statement

The studies involving humans were approved by University of Utah Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

ND: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing. MR: Data curation, Formal Analysis, Software, Writing – review & editing. WL: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – review & editing.

Funding

The authors declare financial support was received for the research, and/or publication of this article. Research reported in this publication was supported by the National Institute of Mental Health under award number R00MH118359 to NdL.

Acknowledgments

Research reported in this publication was supported by the of the National Institutes of Health under award number R00MH118359 to NdL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The support and resources from the Center for High Performance Computing at the University of Utah are also gratefully acknowledged. The computational resources used were partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1. Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development℠ (ABCD) Study (https://abcdstudy.org), held in the NBDC Data Hub. This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. Additional support for this work was made possible from NIEHS R01-ES032295 and R01-ES031074. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from 10.15154/1523041. DOIs can be found at https://nda.nih.gov/abcd/abcd-annual-releases.html.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2025.1487894/full#supplementary-material

References

1. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. *Lancet*. (2007) 369:1302–13.

2. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen HU. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *Int J Methods Psychiatr Res*. (2012) 21:169–84.

3. Pedersen CB, Mors O, Bertelsen A, Waltoft BL, Agerbo E, McGrath JJ, et al. A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry*. (2014) 71:573–81.

4. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. (2005) 62:593–602.

5. Vesterling C, Schütz-Wilke J, Bäker N, Bolz T, Eilts J, Koglin U, et al. Epidemiology of somatoform symptoms and disorders in childhood and adolescence: A systematic review and meta-analysis. *Health Soc Care Community*. (2023) 2023:1–16.

6. Campo JV, Fritsch SL. Somatization in children and adolescents. *J Am Acad Child Adolesc Psychiatry*. (1994) 33:1223–35.

7. Kessler RC, Heeringa S, Lakoma MD, Petukhova M, Rupp AE, Schoenbaum M, et al. Individual and societal effects of mental disorders on earnings in the United States: results from the national comorbidity survey replication. *Am J Psychiatry*. (2008) 165:703–11.

8. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. *Lancet Psychiatry*. (2016) 3:171–8.

9. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*. (2013) 382:1575–86.

10. Roehrig C. Mental disorders top the list of the most costly conditions in the United States: $201 billion. *Health Aff (Millwood)*. (2016) 35:1130–5.

11. Csillag C, Nordentoft M, Mizuno M, McDaid D, Arango C, Smith J, et al. Early intervention in psychosis: From clinical intervention to health system implementation. *Early Interv Psychiatry*. (2017) 12(4):757–64.

12. Conus P, Macneil C, McGorry PD. Public health significance of bipolar disorder: implications for early intervention and prevention. *Bipolar Disord*. (2014) 16:548–56.

13. Hamilton MP, Hetrick SE, Mihalopoulos C, Baker D, Browne V, Chanen AM, et al. Identifying attributes of care that may improve cost-effectiveness in the youth mental health service system. *Med J Aust*. (2017) 207:S27–37.

14. Haque UM, Kabir E, Khanam R. Detection of child depression using machine learning methods. *PloS One*. (2021) 16:e0261131.

15. Huang Y, Zhu C, Feng Y, Ji Y, Song J, Wang K, et al. Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: A cross-sectional study. *J Affect Disord*. (2022) 319:221–8.

16. Garcia-Argibay M, Zhang-James Y, Cortese S, Lichtenstein P, Larsson H, Faraone SV. Predicting childhood and adolescent attention-deficit/hyperactivity disorder onset: a nationwide deep learning approach. *Mol Psychiatry*. (2022) 28 (3):1232–9.

17. Ter-Minassian L, Viani N, Wickersham A, Cross L, Stewart R, Velupillai S, et al. Assessing machine learning for fair prediction of ADHD in school pupils using a retrospective cohort study of linked education and healthcare data. *BMJ Open*. (2022) 12:e058058.

18. Toenders YJ, Kottaram A, Dinga R, Davey CG, Banaschewski T, Bokde ALW, et al. Predicting depression onset in young people based on clinical, cognitive, environmental, and neurobiological data. *Biol Psychiatry Cognit Neurosci Neuroimaging*. (2022) 7:376–84.

19. Chavanne AV, Paillere Martinot ML, Penttila J, Grimmer Y, Conrod P, Stringaris A, et al. Anxiety onset in adolescents: a machine-learning prediction. *Mol Psychiatry*. (2022) 28(2):639–46.

20. Zhang-James Y, Helminen EC, Liu J, Group E-AW, Franke B, Hoogman M, et al. Evidence for similar structural brain anomalies in youth and adult attention-deficit/hyperactivity disorder: a machine learning analysis. *Transl Psychiatry*. (2021) 11:82.

21. Alexander LM, Escalera J, Ai L, Andreotti C, Febre K, Mangone A, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data*. (2017) 4:170181.

22. Bjork JM, Straub LK, Provost RG, Neale MC. The ABCD study of neurodevelopment: Identifying neurocircuit targets for prevention and treatment of adolescent substance abuse. *Curr Treat Options Psychiatry*. (2017) 4:196–209.

23. Karcher NR, Barch DM. The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*. (2021) 46:131–42.

24. Chan L, Simmons C, Tillem S, Conley M, Brazil IA, Baskin-Sommers A. Classifying conduct disorder using a biopsychosocial model and machine learning method. *Biol Psychiatry Cognit Neurosci Neuroimaging*. (2022) 8(6):599–608.

25. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*. (2020) 77:534–40.

26. Bronfenbrenner U. *The ecology of human developments: experiments by nature and design*. Cambridge, MA: Harvard University Press (1979).

27. Cicchetti D, Rogosch FA. Equifinality and multifinality in developmental psychopathology. *Dev Psychopathology*. (1996) 8:597–600. doi: 10.1017/S0954579400007318

28. Sameroff A. The transactional model. In: Sameroff A, editor. *The transactional model of development: how children and contexts shape each other*. American Psychological Association, Washington, DC (2009). p. 3–21. doi: 10.1037/11877-001

29. Dahl RE. The regulation of sleep and arousal: development and psychopathology. *Dev Psychopathology*. (1996) 8:3–27. doi: 10.1017/S0954579400006945

30. Harvey AG. Sleep and circadian functioning: critical mechanisms in the mood disorders? *Annu Rev Clin Psychol*. (2011) 7:297–319. doi: 10.1146/annurev-clinpsy-032210-104550

31. Alfano CA, Gamble AL. The role of sleep in childhood psychiatric disorders. *Child Youth Care Forum*. (2009) 38:327–40. doi: 10.1007/s10566-009-9081-y

32. Ripley BD. *Pattern recognition and neural networks*. Cambridge: Cambridge University Press (1996).

33. Russell S, Norvig P. *Artificial intelligence: A modern approach*. *3rd*. Upper Saddle River, NJ: Pearson Education (2010).

34. Kuhn M, Johnson K. *Applied predictive modeling*. Springer New York, NY: Springer (2013).

35. Jernigan TL, Brown SA, Dowling GJ. The adolescent brain cognitive development study. *J Res Adolesc*. (2018) 28:154–6.

36. Garavan H, Bartsch H, Conway K, Decastro A, Goldstein RZ, Heeringa S, et al. Recruiting the ABCD sample: Design considerations and procedures. *Dev Cognit Neurosci*. (2018) 32:16–22.

37. Volkow ND, Koob GF, Croyle RT, Bianchi DW, Gordon JA, Koroshetz WJ, et al. The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cognit Neurosci*. (2018) 32:4–7.

38. Barch DM, Albaugh MD, Avenevoli S, Chang L, Clark DB, Glantz MD, et al. Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description. *Dev Cognit Neurosci*. (2018) 32:55–66.

39. Lisdahl KM, Sher KJ, Conway KP, Gonzalez R, Feldstein Ewing SW, Nixon SJ, et al. Adolescent brain cognitive development (ABCD) study: Overview of substance use assessment methods. *Dev Cognit Neurosci*. (2018) 32:80–96.

40. Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, et al. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cognit Neurosci*. (2018) 32:43–54.

41. Hagler DJ Jr., Hatton S, Cornejo MD, Makowski C, Fair DA, Dick AS, et al. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage*. (2019) 202:116091.

42. McConaughy SH. The achenbach system of empirically based assessment. In: Andrews JJW, Saklofske DH, Janzen HL, editors. *Educational psychology, handbook of psychoeducational assessment*. San Diego, CA, USA: Academic Press (2001). p. 289–324.

43. Jager S, Allhorn A, Biessmann F. A benchmark for data imputation methods. *Front Big Data*. (2021) 4:693674.

44. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – A practical guide with flowcharts. *BMC Med Res Methodology*. (2017) 17:162. doi: 10.1186/s12874-017-0442-1

45. Novotny PJ, Schroeder D, Sloan JA, Mazza GL, Williams D, Bradley D, et al. Do missing values influence outcomes in a cross-sectional mail survey? *Mayo Clinic Proceedings: Innovations Qual Outcomes*. (2021) 5:84–93. doi: 10.1016/j.mayocpiqo.2020.09.006

46. Dhillon IS, Sra S. Generalized nonnegative matrix approximations with bregman divergences. *Adv Neural Inf Process Syst*. (2006) 18:283–90.

47. Tandon R, Sra S. Sparse nonnegative matrix approximation: new formulations and algorithms. *Max Planck Institute Biol Cybernetics Tech Rep*. (2010) 193:1–17.

48. Xu J, Wang Y, Xu X, Cheng KK, Raftery D, Dong J. NMF-based approach for missing values imputation of mass spectrometry metabolomics data. *Molecules*. (2021) 26:5787.

49. Kursa MB, Jankowski A, Rudnicki WR. Boruta – A system for feature selection. *Fundamenta Informaticae*. (2010) 101:271–85.

50. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. (2013) 36:27–46. doi: 10.1111/j.1600-0587.2012.07348.x

51. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: animperative style, high-performance deep learning library. Advances in neuralinformation processing systems 32. *Curran Associates, Inc*. (2019). 8024–35.

52. Goodfellow I, Bengio Y, Courville A. *Deep learning* Vol. xxii. . Cambridge, Massachusetts: The MIT Press (2016). p. 775.

53. Bishop CM. (2006). *Pattern recognition and machine learning*. New York: Springer.

54. X. B, Varoquaux G. *Survey of machine-learning experimental methods at NeurIPS 2019 and ICLR 2020* (2020). Available online at: https://halarchives-ouvertesfr/hal-02447823 (Accessed September 18, 2025).

55. de Lacy N, Ramshaw M, Kutz JN. *Integrated Evolutionary Learning: an artificial intelligence approach to joint learning of features and hyperparameters for optimized, explainable machine learning*. Front Artificial Intelligence (2022).

56. Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. *Advances in neural information processing systems 30*. Red Hook, NY, USA: Curran Associates, Inc (2017). p. 4765–74.

57. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for pyTorch. *arXiv*. (2020). doi: 10.48550/arXiv.2009.07896

58. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv*. (2017). doi: 10.48550/arXiv.1711.06104

59. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. *arXiv*. (2017). doi: 10.48550/arXiv.1706.03825

60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.5555/1953048.2078195

61. Xiang Q, Chen K, Peng L, Luo J, Jiang J, Chen Y, et al. Prediction of the trajectories of depressive symptoms among children in the adolescent brain cognitive development (ABCD) study using machine learning approach. *J Affect Disord*. (2022) 310:162–71.

62. Owens JA, Spirito A, McGuinn M, Nobile C. Sleep habits and sleep disturbance in elementary school-aged children. *J Dev Behav Pediatr*. (2000) 21:27–36.

63. Alfano CA. (Re)Conceptualizing sleep among children with anxiety disorders: where to next? *Clin Child Fam Psychol Rev*. (2018) 21:482–99.

64. Ranum BM, Wichstrom L, Pallesen S, Falch-Madsen J, Halse M, Steinsbekk S. Association between objectively measured sleep duration and symptoms of psychiatric disorders in middle childhood. *JAMA Netw Open*. (2019) 2:e1918281.

65. Quach JL, Nguyen CD, Williams KE, Sciberras E. Bidirectional associations between child sleep problems and internalizing and externalizing difficulties from preschool to early adolescence. *JAMA Pediatr*. (2018) 172:e174363.

66. Williamson AA, Zendarski N, Lange K, Quach J, Molloy C, Clifford SA, et al. Sleep problems, internalizing and externalizing symptoms, and domains of health-related quality of life: bidirectional associations from early childhood to early adolescence. *Sleep*. (2021) 44.

67. Gregory AM, Rijsdijk FV, Lau JY, Dahl RE, Eley TC. The direction of longitudinal associations between sleep problems and depression symptoms: a study of twins aged 8 and 10 years. *Sleep*. (2009) 32:189–99.

68. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*. (2022) 603:654–60.

69. de Lacy N, Ramshaw MJ. Selectively predicting the onset of ADHD, oppositional defiant disorder, and conduct disorder in early adolescence with high accuracy. *Front Psychiatry*. (2023) 14:1280326. doi: 10.3389/fpsyt.2023.1280326

70. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. *Lancet*. (2012) 379:1056–67.

71. Angold A, Costello EJ, Erkanli A, Worthman CM. Pubertal changes in hormone levels and depression in girls. *Psychol Med*. (1999) 29:1043–53.

72. Brown J, Cohen P, Johnson JG, Smailes EM. Childhood abuse and neglect: specificity of effects on adolescent and young adult depression and suicidality. *J Am Acad Child Adolesc Psychiatry*. (1999) 38:1490–6.

73. Lovato N, Gradisar M. A meta-analysis and model of the relationship between sleep and depression in adolescents: recommendations for future research and clinical practice. *Sleep Med Rev*. (2014) 18:521–9.

74. Pine DS, Cohen P, Brook J. Adolescent fears as predictors of depression. *Biol Psychiatry*. (2001) 50:721–4.

75. Warner V, Weissman MM, Mufson L, Wickramaratne PJ. Grandparents, parents, and grandchildren at high risk for depression: a three-generation study. *J Am Acad Child Adolesc Psychiatry*. (1999) 38:289–96.

76. Franzen PL, Buysse DJ. Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications. *Dialogues Clin Neurosci*. (2008) 10:473–81.

77. Hariri AR, Mattay VS, Tessitore A, Kolachana B, Fera F, Goldman D, et al. Serotonin transporter genetic variation and the response of the human amygdala. *Science*. (2002) 297:400–3.

78. Toenders YJ, van Velzen LS, Heideman IZ, Harrison BJ, Davey CG, Schmaal L. Neuroimaging predictors of onset and course of depression in childhood and adolescence: A systematic review of longitudinal studies. *Dev Cognit Neurosci*. (2019) 39:100700.

79. Whittle S, Lichter R, Dennison M, Vijayakumar N, Schwartz O, Byrne ML, et al. Structural brain development and depression onset during adolescence: a prospective longitudinal study. *Am J Psychiatry*. (2014) 171:564–71.

80. Whittle S, Yap MB, Sheeber L, Dudgeon P, Yucel M, Pantelis C, et al. Hippocampal volume and sensitivity to maternal aggressive behavior: a prospective study of adolescent depressive symptoms. *Dev Psychopathol*. (2011) 23:115–29.

81. Belden AC, Barch DM, Oakberg TJ, April LM, Harms MP, Botteron KN, et al. Anterior insula volume and guilt: neurobehavioral markers of recurrence after early childhood major depressive disorder. *JAMA Psychiatry*. (2015) 72:40–8.

82. Luby JL, Agrawal A, Belden A, Whalen D, Tillman R, Barch DM. Developmental trajectories of the orbitofrontal cortex and anhedonia in middle childhood and risk for substance use in adolescence in a longitudinal sample of depressed and healthy preschoolers. *Am J Psychiatry*. (2018) 175:1010–21.

83. Pagliaccio D, Luby JL, Luking KR, Belden AC, Barch DM. Brain-behavior relationships in the experience and regulation of negative emotion in healthy children: implications for risk for childhood depression. *Dev Psychopathol*. (2014) 26:1289–303.

84. Foland-Ross LC, Sacchet MD, Prasad G, Gilbert B, Thompson PM, Gotlib IH. Cortical thickness predicts the first onset of major depression in adolescence. *Int J Dev Neurosci*. (2015) 46:125–31.

85. Callaghan BL, Dandash O, Simmons JG, Schwartz O, Byrne ML, Sheeber L, et al. Amygdala resting connectivity mediates association between maternal aggression and adolescent major depression: A 7-year longitudinal study. *J Am Acad Child Adolesc Psychiatry*. (2017) 56:983–91 e3.

86. Pan PM, Sato JR, Salum GA, Rohde LA, Gadelha A, Zugman A, et al. Ventral striatum functional connectivity as a predictor of adolescent depressive disorder in a longitudinal community-based sample. *Am J Psychiatry*. (2017) 174:1112–9.

87. Stringaris A, Vidal-Ribas Belil P, Artiges E, Lemaitre H, Gollier-Briant F, Wolke S, et al. The brain's response to reward anticipation and depression in adolescence: dimensionality, specificity, and longitudinal predictions in a community-based sample. *Am J Psychiatry*. (2015) 172:1215–23.

88. Hanson JL, Hariri AR, Williamson DE. Blunted ventral striatum development in adolescence reflects emotional neglect and predicts depressive symptoms. *Biol Psychiatry*. (2015) 78:598–605.

89. Telzer EH, Fuligni AJ, Lieberman MD, Galvan A. Neural sensitivity to eudaimonic and hedonic rewards differentially predict adolescent depressive symptoms over time. *Proc Natl Acad Sci U S A*. (2014) 111:6600–5.

90. Morgan JK, Olino TM, McMakin DL, Ryan ND, Forbes EE. Neural response to reward as a predictor of increases in depressive symptoms in adolescence. *Neurobiol Dis*. (2013) 52:66–74.

91. Whalley HC, Sussmann JE, Romaniuk L, Stewart T, Papmeyer M, Sprooten E, et al. Prediction of depression in individuals at high familial risk of mood disorders using functional magnetic resonance imaging. *PloS One*. (2013) 8:e57357.

92. Chan SW, Sussmann JE, Romaniuk L, Stewart T, Lawrie SM, Hall J, et al. Deactivation in anterior cingulate cortex during facial processing in young individuals with high familial risk and early development of depression: fMRI findings from the Scottish Bipolar Family Study. *J Child Psychol Psychiatry*. (2016) 57:1277–86.

93. Masten CL, Eisenberger NI, Borofsky LA, McNealy K, Pfeifer JH, Dapretto M. Subgenual anterior cingulate responses to peer rejection: a marker of adolescents' risk for depression. *Dev Psychopathol*. (2011) 23:283–92.

94. Rocha TB, Fisher HL, Caye A, Anselmi L, Arseneault L, Barros FC, et al. Identifying adolescents at risk for depression: A prediction score performance in cohorts based in 3 different continents. *J Am Acad Child Adolesc Psychiatry*. (2021) 60:262–73.

95. Li FF, Chen XL, Zhang YT, Li RT, Li X. The role of prepotent response inhibition and interference control in depression. *Cognit Neuropsychiatry*. (2021) 26:441–54.

96. Kooijmans R, Scheres A, Oosterlaan J. Response inhibition and measures of psychopathology: a dimensional analysis. *Child Neuropsychol*. (2000) 6:175–84.

97. Hasegawa A, Matsumoto N, Yamashita Y, Tanaka K, Kawaguchi J, Yamamoto T. Response inhibition deficits are positively associated with trait rumination, but attentional inhibition deficits are not: aggressive behaviors and interpersonal stressors as mediators. *Psychol Res*. (2022) 86:858–70.

98. Winding TN, Andersen JH. Do negative childhood conditions increase the risk of somatic symptoms in adolescence? - a prospective cohort study. *BMC Public Health*. (2019) 19:828.

99. Eminson DM. Medically unexplained symptoms in children and adolescents. *Clin Psychol Rev*. (2007) 27:855–71.

100. Hoffman R, Bibby H, Bennett D, Klineberg E, Rushworth A, Towns S. Family functioning as a protective factor in treating adolescents with complex medico-psychosocial presentations. *Int J Adolesc Med Health*. (2016) 28:437–44.

101. Rhee H, Holditch-Davis D, Miles MS. Patterns of physical symptoms and relationships with psychosocial factors in adolescents. *Psychosom Med*. (2005) 67:1006–12.

102. Craig TK, Cox AD, Klein K. Intergenerational transmission of somatization behaviour: a study of chronic somatizers and their children. *Psychol Med*. (2002) 32:805–16.

103. Schulte IE, Petermann F. Familial risk factors for the development of somatoform symptoms and disorders in children and adolescents: a systematic review. *Child Psychiatry Hum Dev*. (2011) 42:569–83.

104. Janssens KA, Klis S, Kingma EM, Oldehinkel AJ, Rosmalen JG. Predictors for persistence of functional somatic symptoms in adolescents. *J Pediatr*. (2014) 164:900–5 e2.

105. Kingma EM, Janssens KA, Venema M, Ormel J, de Jonge P, Rosmalen JG. Adolescents with low intelligence are at risk of functional somatic symptoms: the TRAILS study. *J Adolesc Health*. (2011) 49:621–6.

106. Rousseau S, Grietens H, Vanderfaeillie J, Hoppenbrouwers K, Desoete A, Van Leeuwen K. The relation between parenting stress and adolescents' somatisation trajectories: a growth mixture analysis. *J Psychosom Res*. (2014) 77:477–83.

107. Mulvaney S, Lambert EW, Garber J, Walker LS. Trajectories of symptoms and impairment for pediatric patients with functional abdominal pain: a 5-year longitudinal study. *J Am Acad Child Adolesc Psychiatry*. (2006) 45:737–44.

108. Berg N, Nummi T, Bean CG, Westerlund H, Virtanen P, Hammarstrom A. Risk factors in adolescence as predictors of trajectories of somatic symptoms over 27 years. *Eur J Public Health*. (2022) 32:696–702.

109. Kato K, Sullivan PF, Evengard B, Pedersen NL. A population-based twin study of functional somatic syndromes. *Psychol Med*. (2009) 39:497–505.

110. Ziv M, Tomer R, Defrin R, Hendler T. Individual sensitivity to pain expectancy is related to differential activation of the hippocampus and amygdala. *Hum Brain Mapp*. (2010) 31:326–38.

111. Fairhurst M, Wiech K, Dunckley P, Tracey I. Anticipatory brainstem activity predicts neural processing of pain in humans. *Pain*. (2007) 128:101–10.

112. Bingel U, Wanigasekera V, Wiech K, Ni Mhuircheartaigh R, Lee MC, Ploner M, et al. The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. *Sci Transl Med*. (2011) 3:70ra14.

113. Keltner JR, Furst A, Fan C, Redfern R, Inglis B, Fields HL. Isolating the modulatory effect of expectation on pain transmission: a functional magnetic resonance imaging study. *J Neurosci*. (2006) 26:4437–43.

114. Cools R, Calder AJ, Lawrence AD, Clark L, Bullmore E, Robbins TW. Individual differences in threat sensitivity predict serotonergic modulation of amygdala response to fearful faces. *Psychopharmacol (Berl)*. (2005) 180:670–9.

115. Browning M, Holmes EA, Murphy SE, Goodwin GM, Harmer CJ. Lateral prefrontal cortex mediates the cognitive modification of attentional bias. *Biol Psychiatry*. (2010) 67:919–25.

116. Chen JY, Blankstein U, Diamant NE, Davis KD. White matter abnormalities in irritable bowel syndrome and relation to individual factors. *Brain Res*. (2011) 1392:121–31.

117. Otti A, Guendel H, Henningsen P, Zimmer C, Wohlschlaeger AM, Noll-Hussong M. Functional network connectivity of pain-related resting state networks in somatoform pain disorder: an exploratory fMRI study. *J Psychiatry Neurosci*. (2013) 38:57–65.

118. Karibe H, Arakawa R, Tateno A, Mizumura S, Okada T, Ishii T, et al. Regional cerebral blood flow in patients with orally localized somatoform pain disorder: a single photon emission computed tomography study. *Psychiatry Clin Neurosci*. (2010) 64:476–82.

119. Dhamala E, Rong Ooi LQ, Chen J, Ricard JA, Berkeley E, Chopra S, et al. Brain-based predictions of psychiatric illness-linked behaviors across the sexes. *Biol Psychiatry*. (2023) 94(6):479–91.

120. Compton WM, Dowling GJ, Garavan H. Ensuring the best use of data: the adolescent brain cognitive development study. *JAMA Pediatrics*. (2019) 173:809–10. doi: 10.1001/jamapediatrics.2019.2081

121. Calzo JP, Blashill AJ. Child sexual orientation and gender identity in the adolescent brain cognitive development cohort study. *JAMA Pediatrics*. (2018) 172:1090–2. doi: 10.1001/jamapediatrics.2018.2496

122. Rozzell K, Moon DY, Klimek P, Brown T, Blashill AJ. Prevalence of eating disorders among US children aged 9 to 10 years: data from the adolescent brain cognitive development (ABCD) study. *JAMA Pediatrics*. (2019) 173:100–1. doi: 10.1001/jamapediatrics.2018.3678

123. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J Mach Learn Res*. (2017) 18:6765–816.