XGBoosting the Signal

A Machine Learning Approach for Informant-Sensitive Predictions of OCD Symptoms in Children Based on Brain Morphology

Margrete Soya Heimvik

Master of Philosophy in Psychology
Cognitive Neuroscience

Department of Psychology
UNIVERSITY OF OSLO

Credits: 60

Spring 2025

Acknowledgements

I would like to thank my supervisors, Øystein Sørensen and Ina Drabløs, for their invaluable guidance and support throughout this thesis. I am also grateful to Esten Høyland Leonardsen, whose mentorship, though not in an official supervisory role, was equally invaluable. Their collaboration and expertise were essential in shaping the direction and quality of this work. I am especially thankful for the opportunity they gave me to deepen my understanding of cognitive neuroscience and research methodology, which has played a key role in developing and strengthening my passion for machine learning.

I also wish to thank the Department of Psychology at the University of Oslo for providing a supportive academic environment and access to key resources. More importantly, I am thankful to LCBC for not only providing the rich dataset for this thesis, but also for offering me a welcoming research community, a space where I could find encouragement and camaraderie from lab mates throughout the process.

Lastly, I would like to extend my gratitude to my fellow classmates, who supported me by listening, letting me vent, and talk through the thesis on more occasions than I can count. To my friends and family—thank you for riding out the highs and lows with me, for your patience through the cortisol-fueled mood swings, and for offering love and understanding exactly when I needed it most.

Abstract

Author: Margrete Soya Heimvik
Title: XGBoosting the Signal: A Machine Learning Approach for Informant-Sensitive Predictions of OCD Symptoms in Children Based on Brain Morphology
Supervisor: Øystein Sørensen **Co-Supervisor:** Ina Drabløs
Author Statement: This thesis is an independent research project and is not part of a larger coordinated study. I approached my supervisors due to my interest in applying machine learning techniques to mental health data and we developed the research idea. I was granted access to the Adolescent Brain Cognitive Development (ABCD) dataset through collaboration with the Centre for Lifespan Changes in Brain and Cognition (LCBC). The data was collected and preprocessed by the ABCD study. However, all subsequent data cleaning, selection, analysis, modeling, and visualization were independently conducted by me. The manuscript was also independently written by me.

Background: Obsessive-Compulsive Disorder (OCD) is a heterogeneous neuropsychiatric condition, often emerging in childhood, characterized by internalizing symptoms (e.g., obsessions) and externalizing symptoms (e.g., compulsions). Assessing symptoms in children is challenging due to frequent discrepancies between informants, most notably between children and their caregivers, which can obscure internalizing symptoms and complicate diagnostic accuracy. As a developmental disorder, structural MRI (sMRI) offers a valuable window into neurodevelopmental trajectories, with emerging evidence suggesting that alterations in brain structure may evolve over time and reflect symptom severity. In light of these challenges, machine learning presents a valuable tool for capturing nuanced patterns in brain structure and symptom expression across informants, potentially improving early identification and the development of personalized, data-driven intervention strategies. *Research Questions:* (1) Can structural brain features predict the severity of OCD-related internalizing symptoms in children? (2) Does predictive accuracy vary systematically between child- and parent-reported symptoms? *Methods*: Parcellated brain volume features were used as input variables in supervised machine learning models developed with the XGBoost algorithm. Internalizing symptoms were assessed using two parallel instruments: the parent-reported Child Behavior Checklist and the child-

reported Brief Problem Monitor, both containing items that support direct informant comparisons. To ensure methodological fairness, identical preprocessing steps, class balancing, and hyperparameter tuning procedures were applied across both models. Performance was evaluated using metrics suited for imbalanced classification. *Results:* Both models showed high classification accuracy; however, this performance was largely driven by class imbalance. Permutation tests indicated that neither model significantly outperformed chance. Although the child-report model displayed slightly better discriminative ability for clinical internalizing symptoms, as indicated by its ROC curve, this did not translate into better classification of clinically significant cases. These results suggest that structural brain features alone provide limited predictive value for internalizing symptom severity. *Conclusion:* This study found little evidence that sMRI-derived structural features can meaningfully predict OCD-related internalizing symptoms in children. The limited results may reflect several factors, including low symptom prevalence, class imbalance, and the relatively coarse resolution of both the imaging and outcome measures. These constraints likely reduced the models' ability to detect clinically meaningful patterns. Future research would benefit from larger, more balanced samples and the incorporation of multimodal data to better capture the complexity of mental health outcomes in children.

Table of Contents

Introduction

Obsessive-Compulsive Disorder (OCD)

Obsessive-compulsive disorder (OCD) is recognized as a prevalent and persistent neuropsychiatric condition, impacting an estimated 2-3% of individuals worldwide (de Mathis et al., 2013). It is defined in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) as the presence of obsessions, compulsions, or both. Obsessions are recurrent, intrusive, and unwanted thoughts, urges, or images that typically cause significant anxiety or distress; they are internal experiences and therefore align with the internalizing symptom dimension. In children, internalizing symptoms such as fear, avoidance, and distress may be especially difficult to detect due to their subjective nature, increasing the likelihood of underdiagnosis. In contrast, compulsions are repetitive behaviors or mental acts performed to reduce anxiety or prevent feared outcomes; they are often outwardly observable and therefore align with the externalizing symptom dimension. Compulsions are typically perceived as excessive and not realistically connected to the outcome.

OCD is documented as a clinically and etiologically heterogeneous disorder, with symptom profiles that vary widely across individuals (Bragdon & Coles, 2017). These symptoms are shaped by a convergence of genetic, neurobiological, and environmental influences (Bragdon & Coles, 2017; Shephard et al., 2021). Onset frequently begins in childhood or adolescence, with up to half of adult cases reporting symptom onset during these stages; prevalence rates in youth ranges from 1-4% (Nazeer et al., 2020). This developmental period is important to consider, as late childhood is characterized by ongoing brain maturation, including early stages of synaptic pruning and myelination in regions implicated in OCD, such as the prefrontal cortex, striatum, and limbic system (Casey et al., 2008). These changes influence emotional regulation and self-control, potentially heightening vulnerability to compulsive and intrusive thought patterns.

Comorbid psychiatric conditions are common in individuals with OCD, further complicating diagnosis and treatment. Studies show that individuals with OCD onset in childhood often have higher rates of co-existing disorders, many of which share features with other diagnosable conditions considered during differential diagnosis (Saad et al., 2017). These

include mood disorders, anxiety disorders, Tourette's disorder, eating disorders, and various personality disorders (Anagnostopoulos et al., 2016; Ivarsson et al., 2008; Wu et al., 2019). Suicidality also emerges as a particularly serious risk in childhood-onset OCD, underscoring the need for comprehensive and accurate assessment strategies (Storch et al., 2017).

One methodological challenge in the study of developmental psychopathology is the frequent discrepancy between informants, particularly between youth and their caregivers, regarding the severity and nature of symptoms. This divergence is well-documented in developmental psychopathology literature and is especially pronounced for internalizing symptoms, which are less observable than externalizing symptoms (Achenbach et al., 1987; De Los Reyes & Kazdin, 2005). Informant discrepancies can lead to significant variations in diagnostic conclusions, treatment planning, and interpretations of clinical or research outcomes. OCD presents a particularly informative case for this investigation, as it encompasses both internalizing and externalizing symptom dimensions, and clinical diagnosis during childhood is often based primarily on parent-reported information.

When studying the symptom dimensions of OCD, traditional statistical models may fall short in handling the high dimensionality and variability inherent in neuroimaging data and symptom reports. To overcome these challenges, machine learning approaches grounded in data-driven pattern recognition and computational modeling offer a promising alternative. These models prioritize predictive accuracy and generalizability, making them well-suited for identifying patterns in noisy data, characterized by high levels of random error or variability, as well as in data compiled from multiple sources (Luxburg & Schoelkopf, 2008; Shmueli, 2011). In recent years, machine learning methods have gained traction in clinical neuroscience, with growing applications in psychiatric diagnosis and biomarker discovery. OCD, with its distinct symptom dimensions, presents an ideal context for examining how brain structure correlates with symptoms reported by different informants.

Neurobiology of OCD

From a neurobiological perspective, OCD has been linked to dysfunction within the cortico-striato-thalamo-cortical (CSTC) circuit, which includes the orbitofrontal cortex, anterior cingulate cortex, thalamus, and basal ganglia (Graybiel & Rauch, 2000). This circuit plays a key role in processes such as cognitive control, error monitoring, and habit formation. Functional

disruptions in CSTC connectivity are thought to underlie core OCD symptoms, including intrusive thoughts and compulsive behaviors.

Structural magnetic resonance imaging (sMRI) studies have repeatedly reported morphological abnormalities in CSTC-related regions in individuals with OCD. Large-scale meta-analytic findings, including those from the ENIGMA-OCD working group, have identified several neuroanatomical alterations in children and adolescents with OCD. These include larger thalamic volumes, reduced cortical thickness in prefrontal and parietal areas, and volumetric changes in subcortical structures such as the caudate, putamen, pallidum, and nucleus accumbens (van den Heuvel et al., 2022; Wang et al., 2022). These structural differences are believed to reflect atypical neurodevelopmental patterns that may impact emotion regulation and behavioral flexibility. These structural abnormalities vary across developmental stages. Children and adolescents with OCD tend to show more pronounced differences in reward-related regions such as the nucleus accumbens and pallidum, whereas adults more commonly exhibit alterations in the amygdala, a region associated with emotional reactivity (Wang et al., 2022). These findings suggest that the neuroanatomical basis of OCD shifts over time, reflecting a developmental progression in the disorder's pathophysiology.

Medication status further influences the interpretation of neuroimaging findings. Studies have shown that unmedicated children with OCD often exhibit larger thalamic volumes compared to healthy controls, a difference not observed in medicated individuals, suggesting that selective serotonin reuptake inhibitors (SSRIs) may normalize or mask structural abnormalities (Wang et al., 2022). SSRIs have also been found to modulate functional activity in CSTC circuits, underscoring their relevance in both structural and functional brain changes (van den Heuvel et al., 2016).

Informant Discrepancies in Child OCD Assessment

Accurately assessing psychological symptoms in children is inherently complex, this is also true for OCD. In both research and clinical settings, evaluations typically depend on multiple informants such as the child, parents, and teachers. However, agreement across these informants is often low to moderate, especially for internalizing symptoms (Achenbach et al., 1987; De Los Reyes & Kazdin, 2005). Studies show that discrepancies between child and parent reports

increase with age and are more pronounced in domains like anxiety and intrusive thoughts (Weisz et al., 2005).

Consider the case of Liam, a 9-year-old who experienced subthreshold symptoms of OCD. For the past year, Liam had been struggling with persistent, intrusive worries about germs, harm coming to loved ones, and making mistakes. These obsessions caused him significant anxiety, yet he engaged in few observable compulsions. His parents, reported minimal concerns, describing Liam as quiet but well-adjusted, with no significant behavioral problems at home. This discrepancy between Liams's self-reported internal distress and his parent's perception of functioning reflects a broader challenge in assessing internalizing symptoms in children and adolescents. This vignette demonstrates the importance of recognizing that the child's perspective is distinct but equally valid. The insights provided by Liam can differ significantly from those of his caretakers, highlighting the potential discrepancies in information regardless of whether the goal is clinical assessment or research.

Despite their challenges, discrepancies are not necessarily errors, they reflect meaningful differences in perspectives, settings, and symptom expressions. The foundational work by Achenbach et al. (1987) emphasizes that "no single informant serves as a gold standard." Rather, multi-informant assessments provide complementary insights and are exceptionally valuable when interpreted with contextual awareness. In OCD, the diagnostic relevance of internalizing symptoms is especially critical. Parent-reported obsessions have been shown to strongly predict categorical OCD diagnoses, reinforcing the diagnostic weight given to caregiver reports in clinical settings (Ivankovic et al., 2024). However, including both child- and parent-report measures of broader internalizing symptoms provide an opportunity to examine informant-specific brain–behavior relationships, a core goal of the present study. Ultimately, understanding how symptom reports diverge, and which brain features correspond to these divergences, may help refine diagnostic tools and promote individualized, informant-sensitive assessment strategies in child and adolescent OCD.

Statistical Learning

Statistical Learning Theory (SLT) underpins many modern machine learning approaches by providing a theoretical framework for learning patterns from data with the goal of making accurate predictions (Luxburg & Schoelkopf, 2008). As a foundational concept in machine

learning, SLT is specifically relevant to supervised learning, a paradigm that trains models on labeled data to map predictors to a response variable. The focus in supervised learning is typically on maximizing predictive accuracy rather than uncovering causal relationship (Shmueli, 2011). Predictive modeling is a core application of supervised learning that involves training probabilistic or algorithmic models to identify patterns in data. These models rely on loss functions, which are mathematical tools used to measure the difference between the model's predicted output and the actual outcome. By minimizing the value of the loss function, the model improves its prediction accuracy and generalizes better to new, unseen data. In clinical neuroscience and neuroimaging, such machine learning methods are increasingly used to associate brain imaging data with predefined categories, such as diagnostic groups. This approach enables the identification of discriminative features that may serve as potential biomarkers for psychiatric or neurological conditions (Enrico et al., 2021).

Learning methods

Linear Models

Linear regressions provide a simple technique for analyzing data by assuming a linear relationship between one or more features (input/predictor) variables (X) and a target (output) variable (Y) and typically estimates parameters using the least squares method. While effective for straightforward linear relationships, they are limited with complex data, which has led to innovative adaptations that offer broader applicability and improved modeling techniques for diverse patterns.

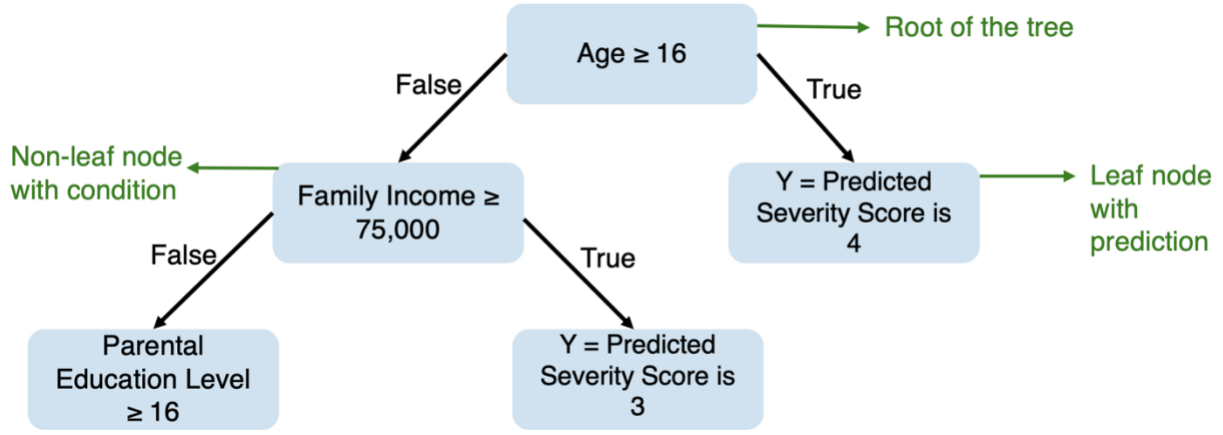
Decision Trees

Decision trees provide a significant advancement to linear models by effectively handling non-linear relationships and interactions between variables. Tree models operate under the assumption that the relationship between the target variable and the features can be captured through locally constant fits, where predictions are based on a fixed value within a small region of the input space rather than a global function like a line (Breiman, 2017). Unlike linear models, which assume a uniform linear relationship across the entire feature space—the set of all possible predictor value combinations—decision trees segment this space using recursive binary splits.

These splits successively divide the data into smaller regions, enabling trees to effectively handle both classification and regression tasks (James et al., 2021). In classification, this results in subsets dominated by a single class, while in regression, it reduces variability in the target values within each subset. As illustrated in Figure 1, each node of the tree serves as a decision point, directing data further down branches or reaching leaf nodes where predictions are determined by metrics such as class majority or mean values. Thus, when used for regression the aim is to split the data into subsets that minimize the resulting mean squared error, mean absolute error, or the variance of the target variable within these subsets (Ryan, 2025). Although they are effective in capturing complex patterns, they are also prone to overfitting, which occurs when the model captures noise, such as random fluctuations or outliers in the training data, rather than the underlying pattern.

Figure 1

Example of a Decision Tree Model Predicting Severity Scores Based on Demographic Features



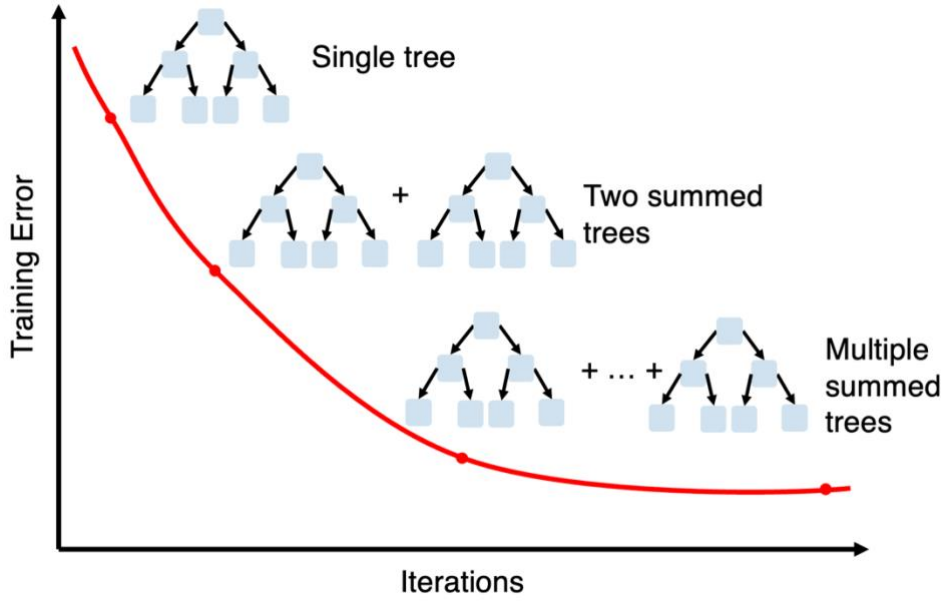
Note: The root node begins with the condition Age ≥ 16. If this condition is met, the model predicts a severity score of 4. If the condition is not met, the decision process continues down to the next node. Each non-leaf node represents a decision based on a feature threshold, while the leaf nodes indicate the predicted severity score (Y). This hierarchical structure illustrates how different combinations of age, income, and education contribute to the final prediction. Adapted from Machine Learning for Tabular Data (1st ed., by M. Ryan & L. Massaron, 2025, Shelter Island, NY: Manning Publications Co. LLC. Copyright 2025 by Manning Publications Co. LLC. Adapted under fair use.

Boosting

Boosting, as illustrated in Figure 2, is an ensemble method used to enhance predictive accuracy (Schapire & Freund, 2012). An ensemble combines multiple models to make more accurate predictions than a single model can produce. Boosting iteratively adds simple models, known as base learners. In this case, the base models are shallow decision trees that improve the overall fit by correcting residuals from previous models (Friedman, 2001). This approach allows diverse loss functions for error minimization, enhancing the alignment of predictions with true outcomes (Ryan, 2025). Adjusting observation weights means assigning more importance to certain training examples. In boosting algorithms, this reweighting strategy is central: after each iteration, the algorithm increases the weights assigned to misclassified or poorly predicted observations, thereby directing subsequent base learners to focus more on these harder-to-predict instances. By doing so, the model iteratively corrects errors while maintaining the overall flexibility and robustness of the ensemble. Flexibility helps to reduce bias by capturing complex patterns, whereas robustness controls variance by preventing overfitting, together supporting an optimal bias-variance tradeoff. XGBoost, or extreme gradient boosting, is a particularly efficient algorithm for fitting boosting models (Ren et al., 2019).

Figure 2

Boosting Process in Ensemble Learning



Note: Training error decreases over successive iterations as additional decision trees are added. Initially, a single tree is trained, followed by subsequent trees that correct the errors of the previous ones. The ensemble prediction is formed by summing the outputs of multiple trees, leading to improved accuracy and reduced training error over time. Adapted from Machine Learning for Tabular Data (1st ed., by M. Ryan & L. Massaron, 2025, Shelter Island, NY: Manning Publications Co. LLC. Copyright 2025 by Manning Publications Co. LLC. Adapted under fair use.

Tuning Parameters

When applying XGBoost to multiclass classification tasks, model performance can be significantly enhanced by fine-tuning key hyperparameters (XGBoost Developers, 2022). While many configurable settings are available, a few are particularly important to highlight due to their influence on the model's ability to distinguish between multiple categories accurately. The learning rate controls how quickly the model learns patterns in the data; smaller values slow learning but often improve generalization. The depth of each tree affects how complex each decision tree is by limiting the number of splits, with deeper trees capturing more intricate patterns but risking overfitting. Additional parameters govern how the training data and features are sampled during each iteration, and how splits are determined based on data distribution and predictive gain. These settings help manage the trade-off between model flexibility and robustness, ensuring that the classifier can generalize well across all classes. Together, these parameters help balance the model's ability to learn complex patterns with its ability to generalize well to new, unseen data.

In conclusion, the application of machine learning techniques, particularly XGBoost, offers a powerful methodology for modeling complex interactions between variables to predict an outcome of interest. This approach provides a robust framework for integrating diverse data types, such as neuroimaging, behavioral, and demographic variables. By using the collective strengths of multiple models, boosted ensembles surpass the predictive capabilities of single decision trees and linear models. This integration of computational tools with clinical insights holds promise for refining diagnostic criteria and enhancing personalized intervention strategies for OCD, paving the way for more precise and effective treatment approaches.

The present study

This thesis is to examines the relationship between structural brain features and OCD-related internalizing symptoms in children. The primary objective of this study was to determine whether brain morphology can reliably predict the severity of internalizing symptoms in children, based on reports from both the children themselves and their parents. In addition to assessing overall model accuracy, the analysis aimed to examine whether predictive performance systematically varied by informant. To that end, a supplementary question was addressed: Is predictive accuracy higher for child-reported internalizing symptoms?

Parcellated brain volume features were extracted from sMRI scans and used as features in supervised machine learning models, implemented using the XGBoost algorithm. The internalizing symptom domain was selected as the predictive target because it is assessed using parallel items across child and parent reports, thus enabling a direct comparison of informant-specific prediction performance. This focus on informant-related variation is motivated by prior findings indicating that internalizing symptoms are more likely to be reported inconsistently and subjectively across sources. By integrating these components, the study seeks to elucidate the relationship between brain structure and subjective symptom expression across informants, thereby contributing to the development of individualized, informant-sensitive predictive models for OCD assessment in child and adolescent populations.

Methods

Data Source and Acquisition

The Adolescent Brain and Cognitive Development (ABCD) Study is a comprehensive decade-long research initiative in the United States designed to enhance our understanding of physical, mental health and risk factors during adolescence (Saragosa-Harris et al., 2022). The study tracks children from ages 9-10 through late adolescence and into early adulthood. This age range covers a crucial developmental stage, where exposure to substance use and the onset of several mental health conditions often take place. The repository includes around 12,000 children at baseline, recruited from 21 research sites (Karcher & Barch, 2021). To ensure the cohort is diverse and representative, the ABCD study employs a multi-stage probability sampling technique to minimize selection bias, thereby enhancing the generalizability of findings generated from the dataset across various studies (Garavan et al., 2018). The ABCD comprises a wide range of behavioral (Barch et al., 2018), multimodal brain imaging (Casey et al., 2018), and other evaluations (Zucker et al., 2018). The data utilized in this thesis is sourced from the ABCD Data Release 5.1 (Haist & Jernigan, 2023).

The ABCD dataset is a publicly available and collected in accordance with ethical standards and approved by institutional review boards (IRBs) at all participating data collection sites. All participants and their legal guardians provided informed consent in accordance with local regulations. The data used in this thesis were accessed through the National Institute of Mental Health Data Archive (NDA) under Approved Data Request #16658, with lead recipient Inge Amlien (LCBC, University of Oslo). Because this project involved the use of anonymized, pre-existing data, no additional approval from the Norwegian Regional Committees for Medical and Health Research Ethics (REK) was required. This thesis utilizes only the year two follow-up time point, as it provided the most comprehensive available data across all key variables of interest, including, neuroimaging data and questionnaire responses.

Structural MRI (sMRI)

The ABCD study collects MRI data from three different scanner platforms located at 21 collection sites across the United States: Siemens Prisma, General Electric (GE) 750, and Philips scanners (Casey et al., 2018). T1-weighted (T1w) images are acquired using a 3D T1w inversion-prepared RF-spoiled gradient echo sequence with 1 mm isotropic resolution (Casey et al., 2018). Prospective motion correction is applied when available (currently only on Siemens and GE scanners; (Tisdall et al., 2012; White et al., 2010)). For the Siemens scanner, acquisition parameters are TR = 2500 ms, TE = 2.88 ms, TI = 1060 ms, flip angle = 8°, with a 256 × 256 matrix, 176 slices, and a 256 mm FOV (acquisition time ~6:08). The Philips scanner used TR = 6.6 ms, TE = 3.1 ms, TI = 950 ms, flip angle = 9°, matrix size 256 × 256, 225 slices, and a FOV of 256 × 240 mm (acquisition time ~5:38). And, the GE scanner, parameters included TR = 2500 ms, TE = 2.0 ms, TI = 1060 ms, flip angle = 8°, with 208 slices and the same resolution and matrix size (acquisition time ~6:09).

The ABCD MRI acquisition protocol utilizes high-density phased array head coils, which can lead to significant variations in image intensity. Additionally, head motion presents a significant challenge since it can degrade image quality and distort derived metrics, this is especially pertinent in adolescent populations where increased movement is more common (Reuter et al., 2015; Satterthwaite et al., 2012). Therefore, although prospective motion correction techniques are implemented to mitigate the effects of motion in sMRI scans, excessive head movement can still introduce substantial artifacts, hindering accurate cortical surface reconstruction and brain segmentation (Tisdall et al., 2016).

Due to the potential artifacts in MRI images, T1w quality control during the MRI acquisition includes three checks (ABCD Study, 2025; Hagler et al., 2019). Firstly, (1) automated checks for protocol compliance assess the completeness of the imaging series and ensure that they meet the specified parameters; these criteria include verifying whether key imaging characteristics such as voxel size and repetition time align with the expected values for each scanner. (2) Automated quality control procedures involve calculating signal-to-noise ratio and head motion statistics. Lastly, this is complemented by (3) a manual quality control process where trained technicians visually assess image quality, identifying and flagging significant artifacts. Series that fail to meet quality standards are excluded from further processing and

analysis, and reviewers are required to document observable artifacts using standardized notations.

Preprocessing sMRI

All sMRI images used in this study were preprocessed and subjected to quality control by the ABCD Study prior to public data release using their standardized in-house image processing pipeline (ABCD Study, 2025). Preprocessing T1w images involves steps to ensure the accuracy and reliability of the data, including (1) Correction for Gradient Nonlinearity Distortions, which addresses distortions in the MRI images introduced by the scanner's gradient system (Jovicich et al., 2006; Wald et al., 2001). These corrections are specific to each scanner model and are provided by MRI manufacturers to enhance image fidelity. (2) Bias Field Correction involves correcting brightness variations across the brain images, a phenomenon known as intensity non-uniformity. This distortion is often caused by the proximity of brain tissue to the MRI coils, leading to areas with extremely high-intensity values that may be erroneously identified as non-brain tissue (i.e., skull). To address this issue, T2-w images are registered to T1-w images using a technique called mutual information, which facilitates accurate alignment and overlay of the different scan types (Wells et al., 1996). Following this registration, the procedure includes tissue segmentation and the application of smoothly varying estimated B1-bias fields to adjust brightness levels, ensuring that each tissue type is represented consistently across the images (Sled et al., 1998). Lastly, (3) Resample to Isotropic: The final image preprocessing step standardizes the viewing and analysis of brain structures. The images are resized and aligned with an internally generated reference brain that features isotropic voxels of 1.0 mm and is approximately aligned with the anterior commissure/posterior commissure (AC/PC) axis (Friston et al., 1995). Moreover, while the ABCD team provides a set of recommended inclusion criteria, accounting for factors such as image quality and protocol compliance, it is the responsibility of the data user to select criteria that are appropriate for their specific analyses. In this thesis, a subset of these criteria was applied based on their relevance to the research objectives.

Brain Segmentation

Cortical surface reconstruction and subcortical segmentation are conducted by the ABCD Study team using FreeSurfer version 7.1.1 (https://surfer.nmr.mgh.harvard.edu). FreeSurfer has been

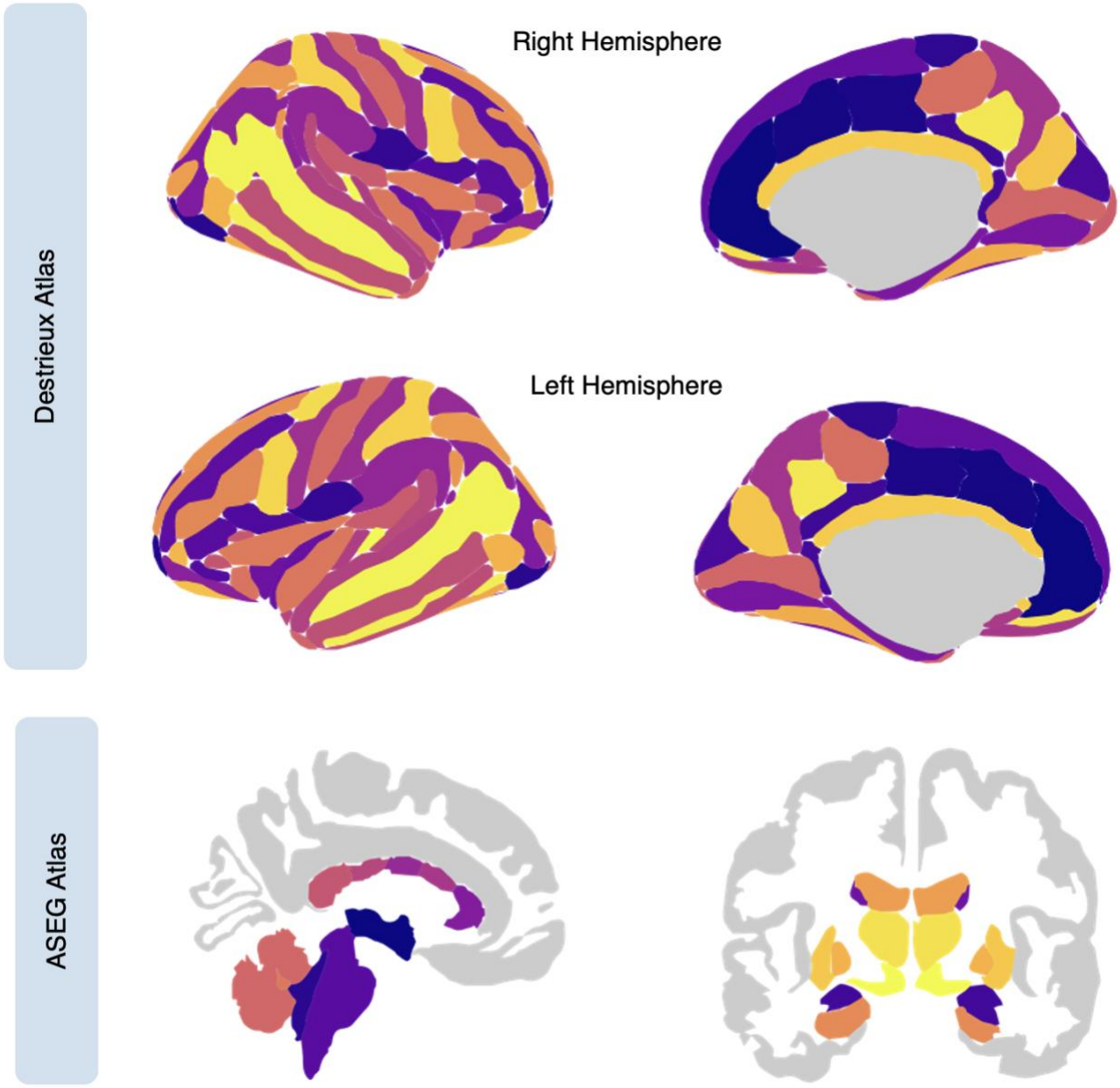
validated in adolescent samples (Biffen et al., 2020). The process begins with skull-stripping, which removes the skull and non-brain tissues from the MRI images (Ségonne et al., 2004). Simultaneously, white matter segmentation is conducted to identify white matter regions, while initial mesh creation produces a preliminary three-dimensional representation of the brain's surface (Dale et al., 1999). Following this, the correction of topological defects on the surface model is applied to address errors or irregularities (Fischl et al., 2001; Segonne et al., 2007). The surface model is optimized and refined (Dale et al., 1999; Dale & Sereno, 1993; Fischl & Dale, 2000). Lastly, the reconstructed brain surface undergoes nonlinear registration to a spherical atlas, aligning it with a standardized spherical model to facilitate consistent comparisons across different subjects (Fischl et al., 1999).

Regions of Interest

After completing cortical reconstruction, specific brain regions are labeled by two atlases. These data are provided by the ABCD Study in tabulated format and made accessible for downstream analyses by independent researchers. Cortical areas are labeled using the Destrieux atlas-based classification (Destrieux et al., 2010). This atlas is widely used in sMRI studies to analyze cortical volume in neurodevelopmental research. This atlas uses a sulco-gyral classification, distinguishing between exposed gyri and buried sulci based on mean curvature and convexity, thus providing 74 bilateral regions (148 total). Subcortical structures are labeled using the Automated Segmentation of the Subcortical Structures (ASEG) provided by FreeSurfer (Fischl et al., 2002). This atlas allows the segmentation and volume measurement of subcortical areas and other intracranial structures, providing 46 regions in total. Combining these atlases facilitates a comprehensive analysis of cortical and subcortical regions. Once both cortical and subcortical structures are labeled, a total of 194 parcellated brain volumes are generated per individual. An overview of the cortical and subcortical parcellation is presented in Figure 3, which illustrates the spatial distribution of labeled regions across both hemispheres and views.

Figure 3

Parcellation of Cortical and Subcortical Brain Structures Using the Destrieux and ASEG Atlases



Note: Visualization of tabulated cortical and subcortical regions was produced using the ggseg packages in R (Mowinckel & Vidal-Piñeiro, 2020). Cortical surfaces (top two rows) are segmented using the Destrieux atlas, which identifies 148 sulcal and gyral regions based on curvature and convexity features. The first row illustrates the right hemisphere in lateral and medial views, while the second row presents the corresponding views of the left hemisphere. Subcortical structures (bottom row) are segmented using FreeSurfer's ASEG atlas, displayed in sagittal (left) and coronal (right) views.

Categorical Psychopathology Assessment

Present psychiatric diagnoses were determined using the Kiddie Schedule for Affective Disorders and Schizophrenia – Computerized Version (KSADS-COMP). It is a self-administered, computerized instrument aligned with DSM-5 diagnostic criteria (J. Kaufman et al., 1997; *KSADS-COMP*, n.d.). The KSADS-COMP evaluates over 50 common childhood and adolescent psychiatric disorders and also provides corresponding ICD-10 codes (Barch et al., 2018). The assessment begins with a structured introductory interview that collects contextual information on family environment, treatment history, gender identity, school functioning, peer relationships (including bullying), and the presence of firearms in the home. This information supports interpretation of mood symptoms and assessment of impairment and risk.

The symptom assessment consists of an initial screening interview, presenting 2–4 key items per disorder. Participants who endorse relevant symptoms are automatically administered full diagnostic supplements through branching logic, ensuring that additional screening is only applied to those for whom it is clinically relevant. Diagnoses are algorithmically assigned as "present," or "not present" based on DSM-5 symptom criteria, duration, and functional impairment. The system also facilitates differential diagnosis, using automated probes to clarify overlapping symptom domains (e.g., mood vs. substance-related conditions). While not administered by a clinician, the KSADS-COMP is clinically designed and validated for use in large-scale studies, offering reliable and scalable mental health phenotyping (J. Kaufman et al., 2017). Demographics, including age, sex, race, and ethnicity, are also retrieved from the KSADS-COMP.

Dimensional Psychopathology Assessment

The Achenbach System of Empirically Based Assessment (ASEBA) is a comprehensive evaluation tool that captures continuous symptom severity and behavioral, emotional and social functioning across various domains (*ASEBA*, 2019). ASEBA is widely applied in diverse areas such as mental health services, education, healthcare, research, and more. The Child Behavior Checklist (CBCL) and Brief Problem Monitor (BPM), two components of the ASEBA, provides a dimensional assessment approach that places behaviors along a continuum of frequency and/or

severity. Raw scores from these instruments are converted into standardized scores using ASEBA-defined algorithms, which account for informant type, age, sex, and ethnicity (Achenbach et al., n.d.). These normed scores are expressed as T-scores, with a mean of 50 and a standard deviation of 10.

Parent-Reported Child Behavior Checklist

The CBCL is a component of the ASEBA first published in 2001 and is a 112-item parent-reported survey, which uses a 3-point Likert scale for responses: "Very True," "Somewhat True," or "Not True," where parents are asked to rate each item based on their child's behavior "now or within the past six months" (Achenbach, 2001). As depicted in Figure 4, the CBCL consists of several dimensions categorized into Syndrome Scales and DSM-Oriented Scales (American Psychiatric Association, 2013; Nelson et al., 2001). The eight syndrome scales are established through factor analysis. They encompass clusters of common behaviors or symptoms. Furthermore, these scales are grouped into three high-level domains: internalizing, externalizing, and total problems scales. These dimensions offer a detailed assessment of a child's emotional, social, and behavioral functioning, aiding in identifying areas that may benefit from therapeutic or educational interventions. The internalizing problems scale specifically captures emotional difficulties such as anxiety, depression, withdrawal, and somatic complaints, making it a key indicator of inwardly directed psychological distress in children. The CBCL/6–18 shows high reliability ($\alpha = 0.83–0.94$; $r = 0.88–0.92$) and strong clinical validity (Achenbach, 2018).

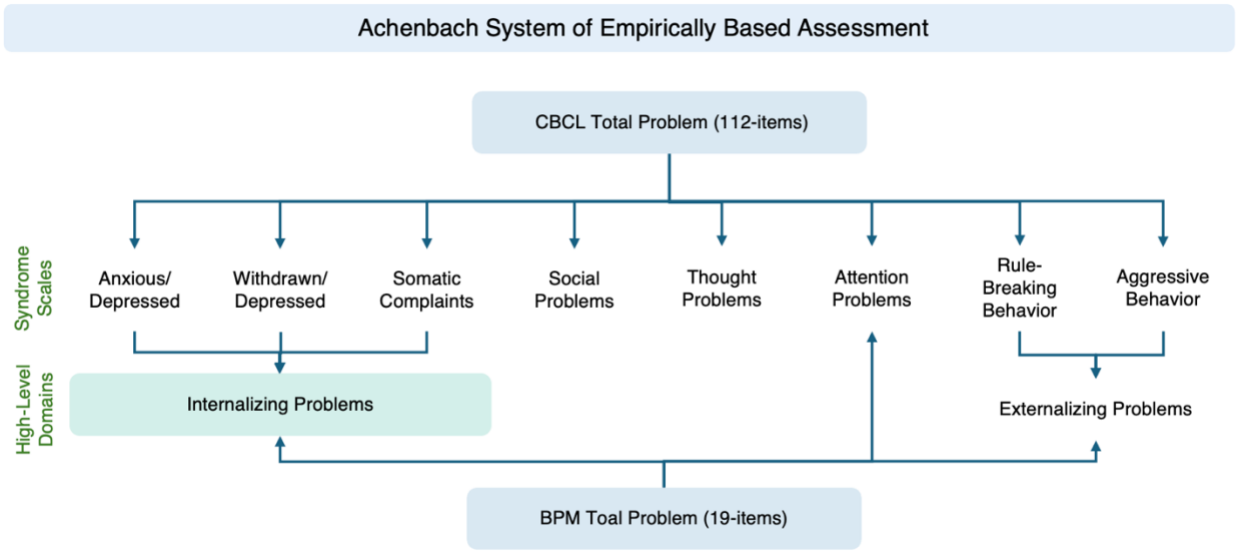
Self-Reported Brief Problem Monitor

The BPM, another component of the ASEBA, was first published in 2011 (Achenbach et al., 2011). Developed to complement parental assessments, adolescents provide self-reports on higher-level domains and attention. It is a 19-item self-reported survey used to assess children's behavioral and emotional functioning and their responses to interventions (RTIs). It also uses a 3-point Likert scale for responses: "Very True", "Somewhat True," or "Not True." Children are instructed to rate each item based on their behavior "currently or within the past six months." As illustrated in Figure 4, the BPM assesses four domains; internalizing, attention Problems, externalizing, and total problems, mirroring the structure of the CBCL/6–18 (Achenbach et al.,

2017). Of particular relevance, the internalizing scale captures symptoms of anxiety, depression, and withdrawal. Recent validation in a Norwegian sample confirmed good internal consistency for this domain (α = .76–.88) and supported the original three-factor structure, affirming its utility as a brief and valid tool for identifying internalizing problems in at-risk children (Pedersen et al., 2021).

Figure 4

The Structure of the ASEBA: Specifically Focusing on the Parallel Between CBCL and the BPM



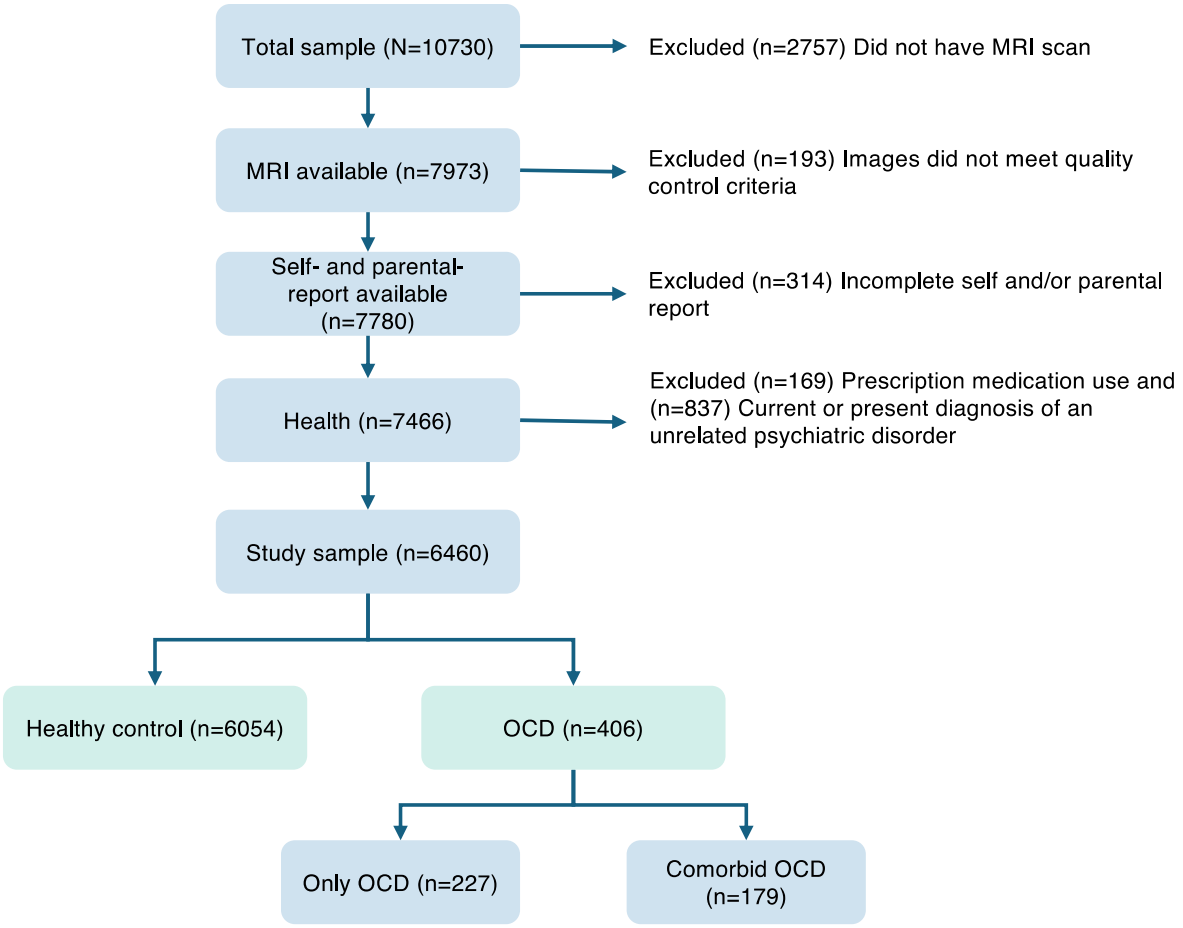
Note: The CBCL consists of Syndrome Scales including clusters of symptoms, which are further grouped into three high-level domains known as (1) internalizing, (2) externalizing, and (3) total problems score that sums all items. The BPM is a shorter version that provides a rapid assessment parallel to dimensions in CBCL for monitoring behavioral and emotional functioning over time. Adapted from "Psychometric Properties of the ASEBA Child Behaviour Checklist and Youth Self-Report in Sub-Saharan Africa—A Systematic Review," by M. R. Zieff, C. Fourie, M. Hoogenhout, and K. A. Donald, 2022, *Acta Neuropsychiatrica*, 34(4), 167–190. https://doi.org/10.1017/neu.2022.5. Copyright 2022 by Cambridge University Press. Adapted under fair use.

Sample

Several criteria were applied to determine the final sample for inclusion in the study. Figure 5 illustrates the participant selection process. Beginning with the full ABCD Year 2 dataset (N = 10,730), participants were excluded if they lacked an MRI scan (n = 2,757) or if their imaging data did not meet quality control standards (n = 193). Of those with usable MRI data, additional exclusions were made for missing self- and/or parent-reported data on internalizing symptoms (n = 314). Moreover, to ensure that medication effects did not confound brain structure findings, participants who were currently prescribed common psychotropic medications (e.g., SSRIs, SNRIs, antipsychotics, stimulants) were excluded (n = 169). Participants who met criteria for a current psychiatric diagnosis unrelated to OCD were excluded from the sample (n = 837). In contrast, individuals with a current OCD diagnosis were retained regardless of comorbid conditions, reflecting the high rates of psychiatric comorbidity commonly observed in OCD (Geller & March, 2012). As a result, healthy control participants were defined as those who did not meet criteria for any current or past psychiatric disorder, based on the KSADS-COMP assessment.

Figure 5

Flowchart of Participant Selection and Subgroup Classification at 2-Year Follow-Up



The final study sample consisted of 6,460 participants, including 6,054 healthy controls and 406 individuals with OCD. Among those with OCD, 227 were classified as having OCD without comorbidities, while 179 presented with at least one comorbid psychiatric condition. Demographic and clinical characteristics for each group are summarized in Table 1, with groups showing comparable distributions across key variables. Among individuals with OCD, the most common comorbid diagnoses were ADHD (n = 71), oppositional defiant/conduct disorder (n = 62), and bipolar disorder (n = 49).

Table 1

Characteristics of The Study Sample by Group

Demographics	**Healthy Control (n=6,054)**	**OCD (n=406)**
Mean Age (SD)	9.47 (0.51)	9.46 (0.50)
Sex		
Female	47,5% (n=2,877)	47.5% (n=193)
Intersex Male	0% (n=1)	-
Male	52.5% (n=3,176)	52.5% (n=213)
Race/Ethnicity		
Asian	2.2% (n=131)	1.0% (n=4)
Black	13.3% (n=806)	15.3% (n=62)
Hispanic	19.2% (n=1,160)	19.5% (n=79)
Other	9.8% (n=593)	14.3% (n=58)
White	55.6% (n=3,364)	50.0% (n=203)
Clinical		
Depressive Disorders	-	n=5
Anxiety Disorders	-	n=46
Attention-Deficit / Hyperactivity Disorder	-	n=71
Oppositional Defiant / Conduct Disorder	-	n=62
Bipolar Disorder	-	n=49
Substance Use Disorder	-	n=2
Suicidality	-	n=39
Eating Disorders	-	n=14

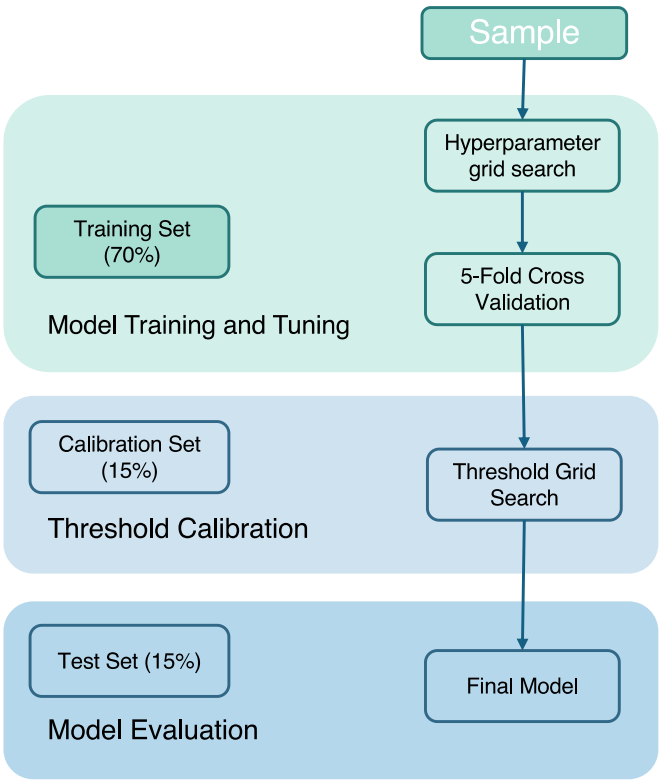
Modelling approach

All analyses were conducted using R Statistical Software (version 4.3.3; R Core Team 2021), and all models were implemented using the XGBoost 3.0 package. Under the XGBoost framework, model performance is optimized by minimizing a specified loss function which

defines the objective of the model. The objective function serves as a core component of model specification, as it directly reflects the type of prediction task being addressed (Brownlee, 2021). To prevent data leakage, the dataset was stratified and split into three subsets prior to any model building or feature selection, ensuring that each set included a similar distribution of symptom severity categories. As illustrated in Figure 6, 70% (n=4522) of the data was allocated for training and tuning the hyperparameters of the model, 15% (n=969) for threshold calibration, and the remaining 15% (n=969) for model evaluation. Full details of the data partitioning are provided in Appendix I. The data split ensured that information from the evaluation set did not inadvertently influence model training or parameter tuning (S. Kaufman et al., 2012).

Figure 6

Flowchart Illustrating the Machine Learning Pipeline



Note: The full sample was split into three subsets. In the training phase, hyperparameter optimization was performed on the training set using grid search and 5-fold cross-validation. In

the threshold calibration phase, a threshold grid search for class-specific probability cutoffs was used. Finally, in the evaluation phase, the calibrated model was applied to a held-out test set to assess final performance metrics.

Exploratory Modeling and Feature Selection

As an initial step, simpler linear regression models were tested to evaluate the predictive utility of different sets of variables. These models included demographic features (age, sex, and race/ethnicity), latent psychosocial factors (socioeconomic status, social risk, and perinatal risk), and sMRI. The goal of these preliminary models was to assess the individual and combined contributions of each feature domain and to establish a baseline for comparison. Full details and model summaries are provided in Appendix II. Subsequent analyses focused exclusively on structural brain features. This decision was guided by the primary aim of the study (to investigate brain-based prediction of OCD-related internalizing symptoms), as well as by practical constraints associated with the high dimensionality of neuroimaging data relative to sample size, which necessitated dimensionality reduction to prevent overfitting.

Initial Regression Approach

The prediction of the internalizing symptoms was initially formulated as a regression task, using the continuous T-score derived from the Achenbach scale for the CBCL and the BPM (Achenbach, 2001; Achenbach et al., 2011). The model objective was set to minimize the squared error loss function. However, the distribution of symptom severity was found to be highly skewed, with relatively few individuals classified within the clinical range. This imbalance in the target variable resulted in suboptimal model performance and limited interpretability, as the squared error loss function assumes equal importance across all prediction errors, irrespective of class prevalence (Hastie et al., 2009). In contrast, classification models are better equipped to handle imbalanced outcome distributions, particularly when combined with class weighting strategies that adjust for disparities in group sizes (He & Garcia, 2009).

Redefining the Task as Multiclass Classification

To improve clinical interpretability and model performance, the task was reframed as a multiclass classification problem by mapping T-scores to established clinical categories (Achenbach, 2009): normal (<65), borderline (65–69), and clinical (≥70). Consequently, the model objective was set to minimize the multiclass log-loss (cross-entropy), which provides a probabilistic output across all categories (*XGBoost*, 2022). This approach, mathematically equivalent to multinomial logistic regression, enables a symmetric and interpretable estimation of class membership probabilities (James et al., 2021). Multiclass log-loss is particularly suited for imbalanced classification settings, as it penalizes confident but incorrect predictions more heavily than uncertain ones. This results in a more informative loss signal than traditional accuracy metrics, which may obscure poor performance on minority classes (Niculescu-Mizil & Caruana, 2005).

Addressing Class Imbalance with Cost-Sensitive Learning

To enhance classification performance in the context of unequal class distributions, both data-level and algorithm-level strategies were employed. A commonly used approach for addressing class imbalance involves assigning greater weights to minority class samples during model training, thereby mitigating the tendency of the model to prioritize the majority class (Kuhn & Johnson, 2013; Ting, 2002). In this study, class weights were applied to increase the influence of participants categorized in the borderline and clinical symptom groups, relative to those in the normal range. This strategy was particularly important given the disproportionately smaller size of the clinically significant group and the high dimensionality of the neuroimaging feature space, both of which increase the risk of model bias, variance inflation, and overfitting. Weighting the minority classes ensured that the model remained sensitive to clinically meaningful patterns, even when those patterns were underrepresented in the training data.

Hyperparameter Tuning and Cross Validation

Hyperparameter optimization was conducted using a grid search across a predefined set of key tuning parameters, including maximum tree depth, learning rate, regularization strength, and row and column sampling ratios (Xgboost Grid Search - R, n.d.). To evaluate model performance of

the tuned models, a five-fold cross-validation procedure was implemented on the training dataset, with early stopping. Early stopping is a regularization technique that terminates model training when performance on a validation set ceases to improve after 10 iterations, thereby preventing overfitting and reducing training time (Prechelt, 1998). The grid search explored 864 unique parameter combinations, each evaluated across five validation folds, resulting in a total of 4,320 model fits. The configuration that achieved the lowest average multiclass log-loss across validation folds was selected as the final model. This tailored tuning procedure enabled the application of stronger regularization, controlled model complexity, and optimized sampling strategies, collectively contributing to a more robust, stable, and generalizable classifier suitable for high-dimensional neuroimaging data (James et al., 2021; Kuhn & Johnson, 2013).

Threshold Calibration

To improve classification performance and ensure that predicted labels aligned more closely with clinically meaningful groupings, a post-hoc threshold calibration was performed using the class probabilities generated by the XGBoost model (*3.3. Tuning the Decision Threshold for Class Prediction*, n.d.). In standard multiclass classification, labels are assigned using the argmax rule, which selects the class with the highest predicted probability. However, this default strategy can introduce bias toward the majority class, particularly in imbalanced datasets, which are common in clinical research contexts (Van Calster et al., 2019). This bias is rooted in the objective function of most machine learning classifiers, including XGBoost, which typically aim to minimize overall loss (e.g., log-loss) across all samples (He & Garcia, 2009). As a result, the default argmax rule systematically favors majority class predictions, even when the minority class probabilities are clinically meaningful. To address this issue, a class-specific threshold calibration approach was implemented. Using a one-vs-rest (OvR) framework on a held-out validation set, the model evaluated the decision threshold for each class independently (Rifkin & Klautau, 2004). Accordingly, each category—normal, borderline, and clinical—was assessed independently during threshold calibration. For each binary classification (e.g., Class A vs. not Class A), the optimal decision threshold was determined by maximizing balanced accuracy, which accounts for both sensitivity (true positive rate) and specificity (true negative rate), providing a more equitable measure of performance in imbalanced datasets (Brodersen et al.,

2010). This process yielded individualized decision thresholds for each class, providing a more flexible alternative to the standard argmax-based decision rule.

Evaluation Metrics

The final tuned and calibrated models' performance was assessed on a held-out test set. The performance was assessed using a comprehensive set of evaluation metrics selected to address the imbalanced nature of the outcome classes, particularly the underrepresentation of individuals with clinically significant OCD symptoms. Overall accuracy was not used as the sole performance indicator, as it can obscure poor detection of minority classes in imbalanced datasets (Saito & Rehmsmeier, 2015).

Instead, a multidimensional evaluation strategy was implemented. The caret package was used to compute confusion matrices and extract key metrics, including balanced accuracy, sensitivity, specificity, and positive predictive value for each class (Kuhn, 2008). Balanced accuracy was used to account for class imbalance by averaging sensitivity (true positive rate) and specificity (true negative rate) for each class, providing a more equitable evaluation of performance (Brodersen et al., 2010). Precision, recall, and F1 scores were calculated for each class using the MLmetrics package (Yan, 2016). Precision reflects the proportion of true positives among all positive predictions, while recall captures the proportion of actual positives correctly identified. The F1 score, defined as the harmonic mean of precision and recall, offers a balanced measure of classification performance that is particularly valuable when the costs of false positives and false negatives are asymmetric. To assess the model's ability to discriminate between classes, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed using a OvR approach with the pROC package (Robin et al., 2011). These metrics visualize and quantify the trade-off between sensitivity and specificity across thresholds and offer insight into the model's separability across all classes.

Post-Hoc Externalizing Analysis

As an exploratory analysis, predictive performance was also examined for parent-reported externalizing symptoms. Results for this analysis are presented in Appendix III.

Permutation Based Significance Testing

To evaluate whether the model's performance exceeded what could be expected by chance, a permutation test was conducted following established methods for assessing statistical significance in predictive modeling (Good, 2000; Ojala & Garriga, 2009). This non-parametric approach involves disrupting the relationship between input features and class labels while preserving the underlying feature distributions. For computational efficiency, hyperparameter tuning was performed once using the original labels and fixed for all permutation iterations. This approach may slightly underestimate the variance in the null distribution (Ojala & Garriga, 2009).

To simulate the null hypothesis of no association between features and target labels, class labels in the held-out test set were randomly permuted 1,000 times. For each permutation, the trained XGBoost model was used to generate predicted class probabilities, which were thresholded using the previously calibrated class-specific thresholds. These thresholded outputs were then converted into class predictions. The classification accuracy was computed for each of the 1,000 permutations, producing a null distribution of accuracies expected under chance. The observed accuracy, obtained using the true (non-permuted) labels, was then compared against this null distribution. A permutation-based p-value was calculated as the proportion of permuted accuracies that were greater than or equal to the observed accuracy, providing a robust estimate of statistical significance without assuming distributional properties of the data. A density plot was generated to visualize the null distribution, with the observed accuracy overlaid as a vertical reference line, highlighting its deviation from the permutation-based baseline.

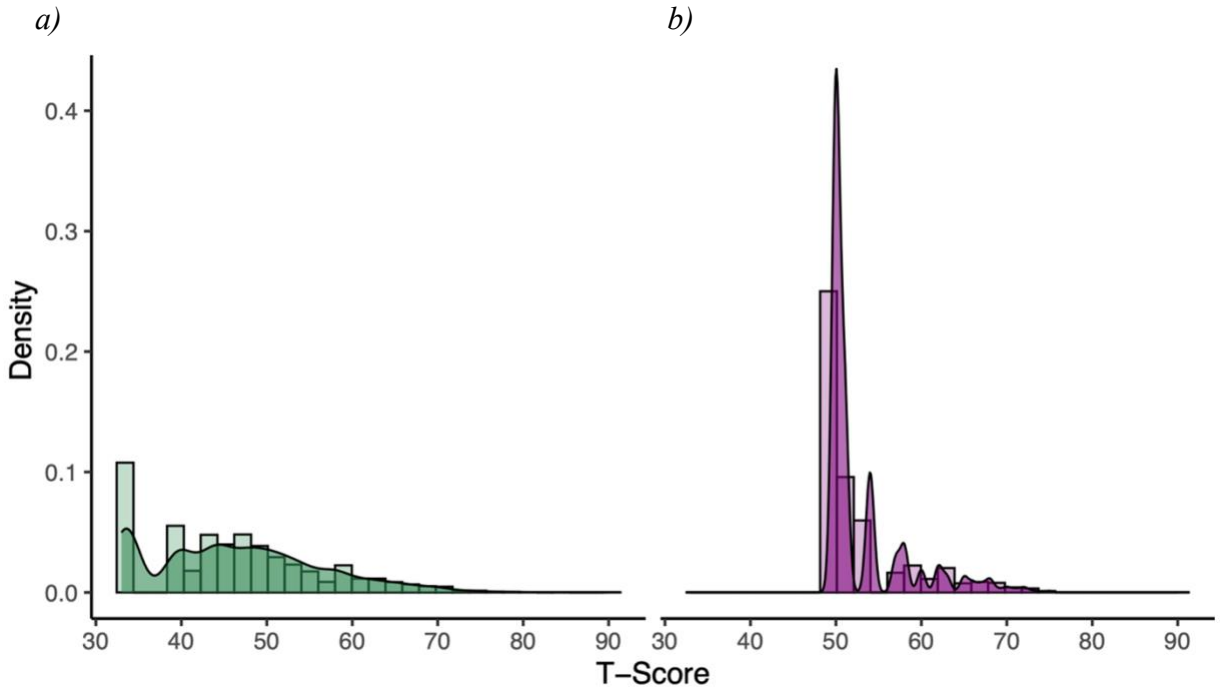
Results

Internalizing Symptom Score Distribution and Baseline Models

The distribution of internalizing symptom scores was examined for both the parent-report and child-report scales, see Figure 7. Both measures exhibited strong positive skew, with the majority of participants falling within the normal range (T < 65) and relatively few classified in the borderline (65–69) or clinical (≥70) ranges. This skew was more pronounced in the child-report distribution, which demonstrated a sharp mode near the normative threshold, highlighting the low prevalence of clinical internalizing symptoms reported in the sample. Child-reported internalizing scores had a mean of 53.10 (SD = 5.13), ranging from 50 to 75. Parent-reported scores had a mean of 46.37 (SD = 9.87), with a wider range of 33 to 90.

Figure 7

Distribution of Target Variable: Internalizing Symptom T-Scores for Parent- and Child-Reported Measures



Note: Density plots showing the distribution of internalizing T-scores (N=6560) derived from the (a) parent-reported (CBCL) in green, and (b) child-reported (BPM) in purple.

As summarized in Table 2, model performance under the default configuration—without class weighting, hyperparameter tuning, or threshold calibration—was near chance level for both models. Overall accuracy was low (CBCL: 33.3%; BPM: 31.4%), closely mirroring the expected performance of a model making random predictions across three categories (healthy, borderline, clinical). Cohen's kappa values were negative or near zero (CBCL: –0.0074; BPM: –0.0136), indicating no meaningful agreement between predicted and actual classifications. Balanced accuracy was similarly poor (CBCL: 49.4%; BPM: 47.5%), and class-level sensitivity for the borderline (CBCL: 28.6%; BPM: 22.7%) and clinical groups (CBCL: 41.7%; BPM: 38.9%) remained limited. These findings suggest that the base models lacked both discriminative ability and clinical utility.

Applying class weights yielded only modest improvements. Accuracy remained nearly identical to the default models, and while Cohen's kappa improved slightly, it still indicated minimal agreement beyond chance. Balanced accuracy increased marginally, with some gains in sensitivity for borderline and clinical cases. However, these changes were insufficient to enable reliable classification of minority classes, suggesting that class reweighting alone could not compensate for the effects of class imbalance or overlapping symptom profiles.

Table 2

Classification Performance Metrics for Parent- and Child-Reported Internalizing Symptoms Across Four XGBoost Configurations

	CBCL				**BPM**			
Model Configuration	Default	With Weights	Tuned + Argmax	Tuned + Threshold-Cal.	Default	With Weights	Tuned + Argmax	Tuned + Threshold-Cal.
Overall Performance								
Accuracy	0.3333	0.3323	0.9505	0.6801	0.3137	0.3158	0.934	0.9247
95% CI	0.3037–0.364	0.3027–0.363	0.9349–0.9633	0.6497-0.7094	0.2846–0.344	0.2866–0.3461	0.9164–0.9488	0.9062–0.9405
Cohen's Kappa	-0.0074	0.0224	-0.0016	-0.0289	-0.0136	-0.009	0.0	-0.0168
Class Balance								
Balanced Accuracy	0.4942	0.5687	0.4997	0.465	0.4745	0.4795	0.5	0.4966
Sensitivity (Borderline)	0.2857	0.4524	0.0	0.048	0.2273	0.2955	0.0	0.0
Sensitivity (Clinical)	0.4166	0.5417	0.0	0.083	0.3889	0.3334	0.0	0.0
Specificity (Borderline)	0.6505	0.6440	0.9989	0.865	0.6605	0.6681	1.0	0.9881
Specificity (Clinical)	0.6878	0.6804	1.0	0.859	0.6572	0.6509	1.0	1.0
Class-Level Discrimination								
Precision (Borderline)	0.0357	0.0544	0.0	0.016	0.0309	0.0407	NaN	0.0
Precision (Clinical)	0.0327	0.0413	NaN	0.015	0.0210	0.0178	NaN	NaN
F1 Score (Borderline)	0.0634	0.0967	0.0	0.0236	0.0538	0.0723	0.0	0.0
F1 Score (Clinical)	0.061	0.0763	0.0	0.0248	0.0382	0.0334	0.0	0.0

Note: Accuracy refers to the overall proportion of correct predictions. 95% CI indicates the range in which the true accuracy likely falls with 95% confidence. Cohen's kappa adjusts accuracy for chance agreement, with values near zero indicating chance-level performance. Balanced accuracy averages sensitivity and specificity across all classes, providing a more informative metric under class imbalance. Sensitivity (also called recall) measures how well the model identifies true positives within each class. Specificity reflects the model's ability to correctly identify true negatives. Precision is the proportion of predicted cases that are actually correct. F1 Score combines precision and sensitivity, offering a balanced measure of class-specific performance, particularly useful when class distributions are skewed. While argmax models showed high overall accuracy, this reflected overprediction of the healthy class. Class-specific metrics revealed poor detection of minority classes, with only modest gains after threshold calibration for parent reports and persistently low performance for child-reports, underscoring the impact of class imbalance and limited signal in the input data. Precision values for some classes are missing (NaN). This occurs because the model did not predict any instances for those classes, resulting in a division by zero when calculating precision. These NaNs highlight a failure to identify minority classes.

Tuning Parameters and Loss Minimization Across The Grid Search

As shown in Table 3, the optimal tuning parameter configuration for both models converged at 50 boosting rounds. While core parameters such as learning rate and max depth were consistent, the parent-report model used a lower minimum child weight (1 vs. 5), higher gamma (2 vs. 0), and differed in subsampling strategies (subsample: 0.6 vs. 0.9; colsample by tree: 0.9 vs. 0.6). This suggests that the parent-report model adopts a more flexible architecture, allowing for finer-grained splits in the data. However, this flexibility is tempered by stronger regularization mechanisms, such as higher gamma values and more restrictive subsampling strategies. These controls help prevent overfitting by limiting unnecessary model complexity. In contrast, the child-report model employs a more conservative structural configuration, imposing stricter thresholds on data partitioning (e.g., higher minimum child weight). Rather than relying on regularization through sampling or split constraints, it restricts model complexity primarily through these foundational parameter settings.

Hyperparameter tuning revealed variation in model performance across the parameter space, see Figure 8. For both the parent-report (Figure 8a) and child-report (Figure 8b) models, log loss values ranged from approximately 1.090 to 1.108, indicating sensitivity to tuning configurations. Performance was most strongly influenced by tree depth and learning rate, with lower log loss achieved under shallower trees and moderate learning rates. Despite the broad search space, both models converged on similar optimal configurations, yielding final multiclass log-loss values of 1.090 (CBCL) and 1.093 (BPM). Class weighting contributed to improved classification of underrepresented borderline and clinical categories (see Table 2).

Table 3

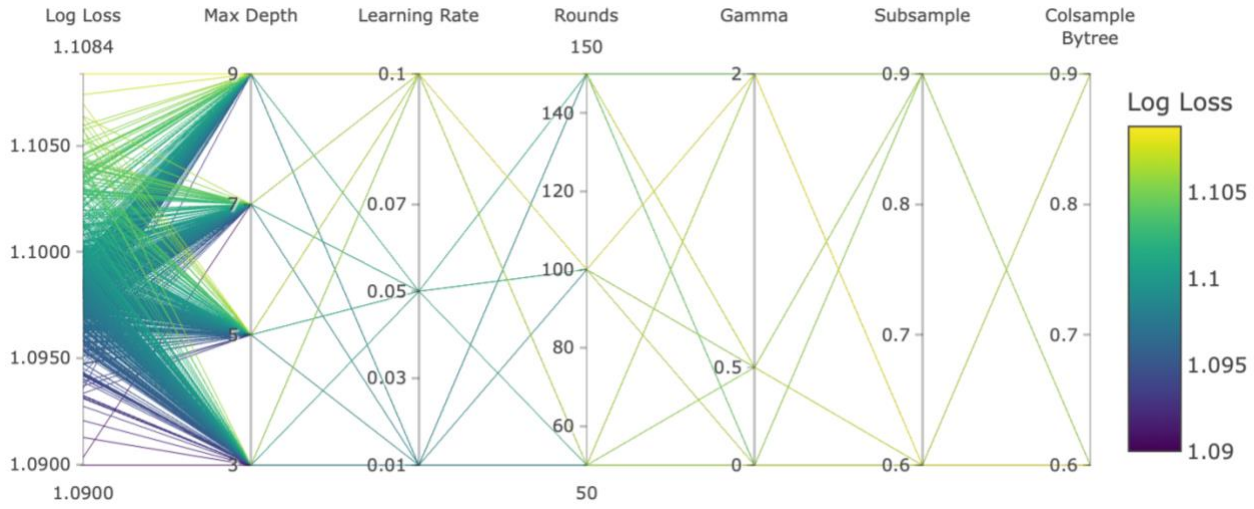
Optimal Tuning Parameter Configuration for Internalizing Symptom Prediction Models

Parameter	**CBCL**	**BPM**
Booster	Gbtree	Gbtree
Objective	Multi:softprob	Multi:softprob
Evaluation Metric	mlogloss	mlogloss
Max Depth	3	3
Min Child Weight	1	5
Eta	0.1	0.1
Gamma	2	0
Subsample	0.6	0.9
Colsample by Tree	0.9	0.6
Number of Classes	3	3
Best number of rounds	50	50
Best multiclass Log Loss	1.090	1.093

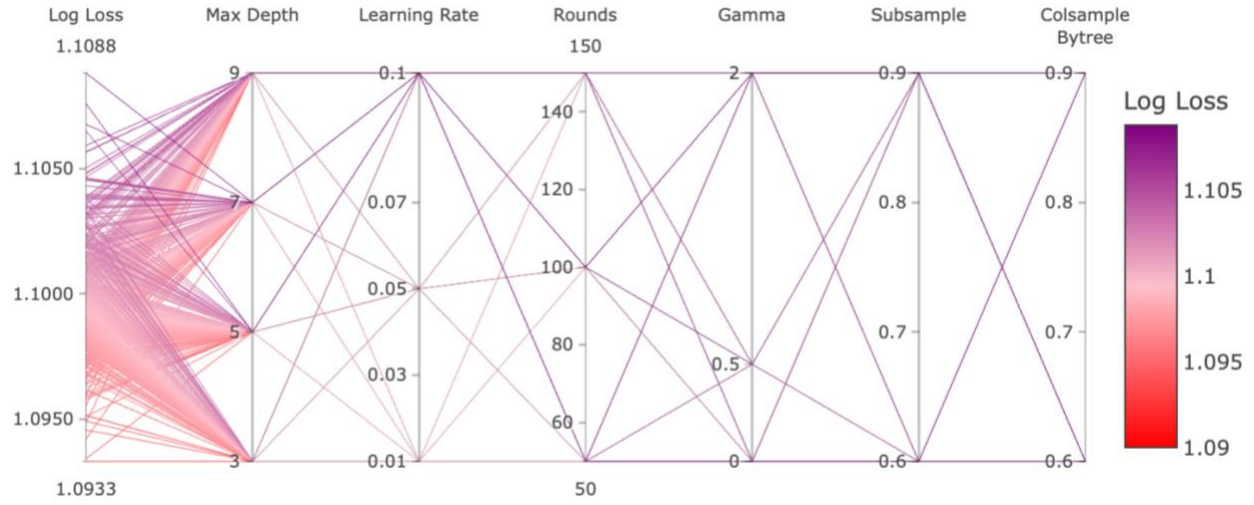
Figure 8

Hyperparameter Optimization Results for Internalizing Symptom Prediction Models

a)



b)



Note: Parallel coordinate plots illustrating how performance varies across the hyperparameter space. Each line represents a unique combination of tuning parameters across the grid search, with axes corresponding to individual hyperparameters and the resulting multiclass log loss on the y-axis. (a) . Parent-reported symptoms (CBCL); line color indicating yellow as lower loss and blue as higher loss. (b) Child-reported symptoms (BPM); line color indicates log loss, with red representing lower loss and purple representing higher loss.

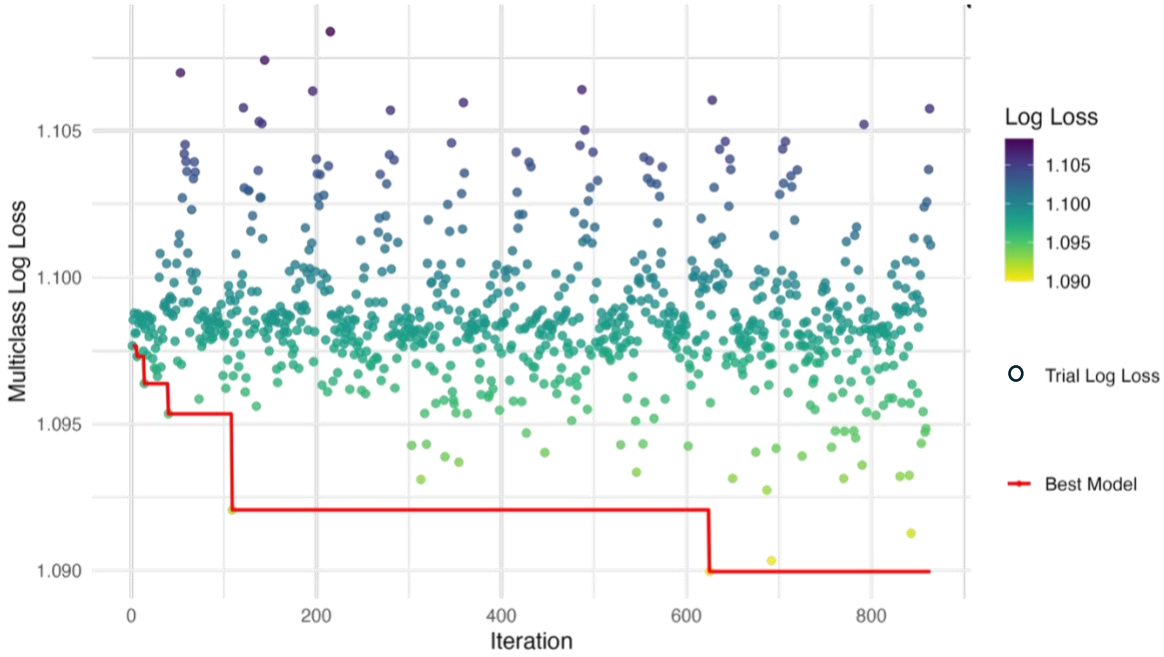
Optimization history across the 864 hyperparameter configurations is visualized in Figure 9. Performance, as indicated by the minimum multiclass log loss, showed distinct patterns across models. The parent-report model (Figure 9a) achieved its lowest log loss earlier in the tuning process, after which subsequent configurations showed diminishing returns. In contrast, the child-report model (Figure 9b) performance improved gradually across the tuning process, with the lowest log loss identified near the end of the search.

Notably, as summarized in Table 2 the hyperparameter-tuned models without threshold calibration yielded substantial increases in overall accuracy (CBCL: 95.05%; BPM: 93.4%). Closer inspection revealed that these accuracy gains were primarily due to the models defaulting to the majority (healthy) class. By consistently predicting the most common label, the models correctly classified most healthy individuals but failed to identify any borderline or clinical cases. Thus, despite the apparent accuracy, Cohen's kappa values were near zero (CBCL: –0.0016; BPM: 0.000), and balanced accuracy remained at chance (CBCL: 0.4997; BPM: 0.5000), reflecting the models' inability to detect elevated symptom profiles despite tuning efforts.

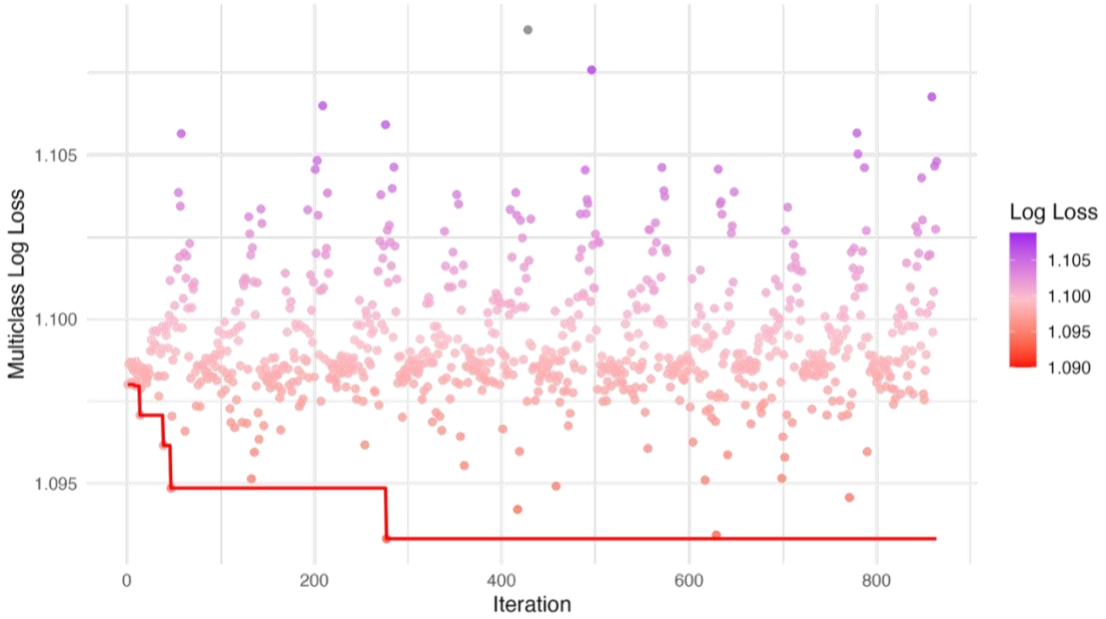
Figure 9

Tuning Parameter Optimization History Across Iterations

a)



b)



Note: Optimization history of multiclass models across hyperparameter configurations. Each point represents a distinct hyperparameter set, with color indicating the corresponding multiclass log loss (lower values shown in red/yellow; higher values in purple). Trial Log Loss refers to the

multiclass log loss obtained from each individual training run using a specific hyperparameter set. The red step line tracks the running minimum log loss, highlighting how progressively better-performing model configurations were identified through the boosting process, which iteratively corrects the errors (residuals) of preceding models. (a) Optimization trajectory for the model predicting parent-reported (CBCL) internalizing symptoms. (b) Optimization trajectory for the model predicting child-reported (BPM) internalizing symptoms.

Threshold Calibration

Threshold calibration using a one-vs-rest strategy was implemented to enhance sensitivity for underrepresented classes. As shown in Table 2, this adjustment led to a decline in overall accuracy (CBCL: 68.0%; BPM: 92.5%) and further reductions in Cohen's kappa (CBCL: –0.0289; BPM: –0.0168). Although calibration redistributed predicted probabilities it did not yield meaningful improvements in minority class detection. Sensitivity for the clinical group remained extremely low (CBCL: 8.3%; BPM: 0.0%), and borderline sensitivity was similarly poor (CBCL: 4.8%; BPM: 0.0%). These findings suggest that threshold calibration, while mitigating some majority-class bias, was insufficient to achieve clinically relevant classification performance, particularly in the child-report model where class imbalance was most severe.

Model Performance and Class Discrimination

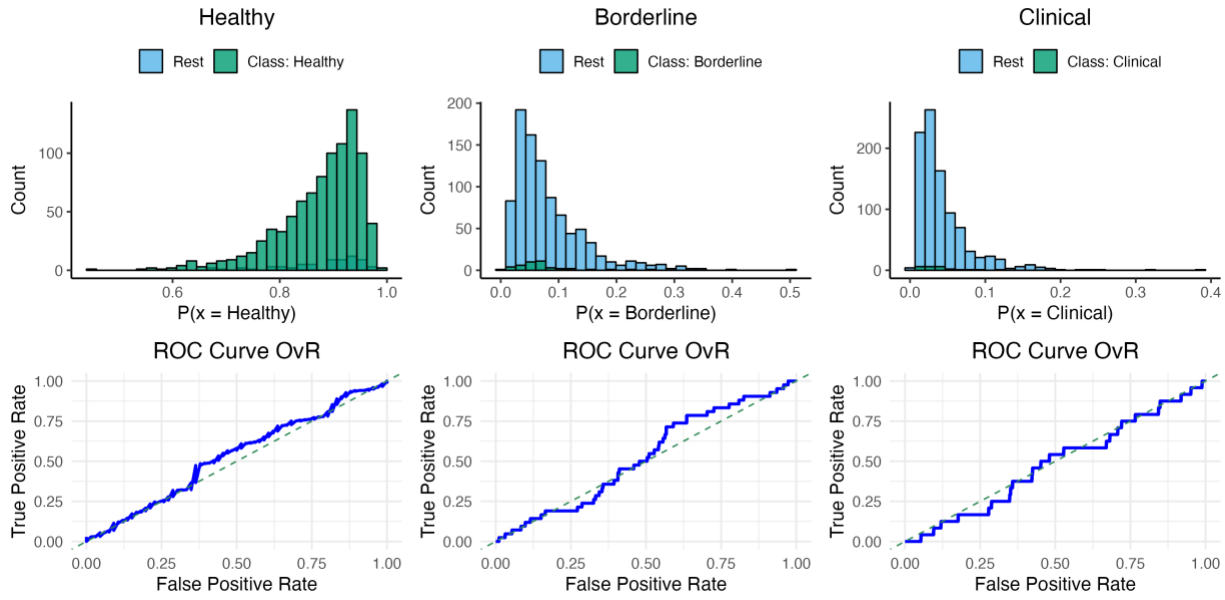
Overall model classification performance and class discrimination ability for both parent- and child-reported internalizing symptoms were limited, even after threshold calibration, as illustrated by ROC curves and class probability distributions, see Figure 10. The parent-report model, Figure 10a, exhibited near-diagonal ROC curves for both borderline and clinical classes and output probabilities that reflected low certainty and poor separation between classes. This indicates near-random classification and low confidence for minority outcomes. The child-report model, Figure 10b, showed some upward curvature for the clinical group ROC curve, suggesting limited discriminatory ability, whereas the borderline group remained poorly differentiated. Predicted probabilities for healthy cases were more confidently distributed near 1.0, while minority class probabilities clustered near zero, reflecting poor model certainty. Overall, these

findings highlight the persistent difficulty of achieving clinically meaningful classification performance in imbalanced symptom categories, particularly for child-reported outcomes.

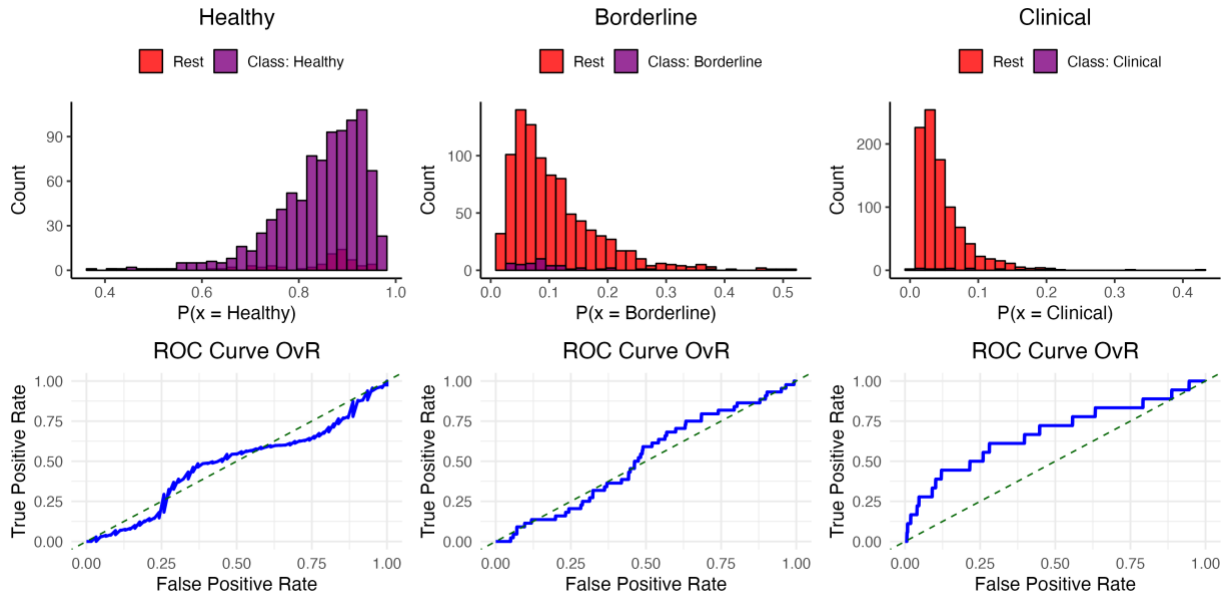
Figure 10

ROC Curves and Class Probability Distributions for the Test Set

a)



b)



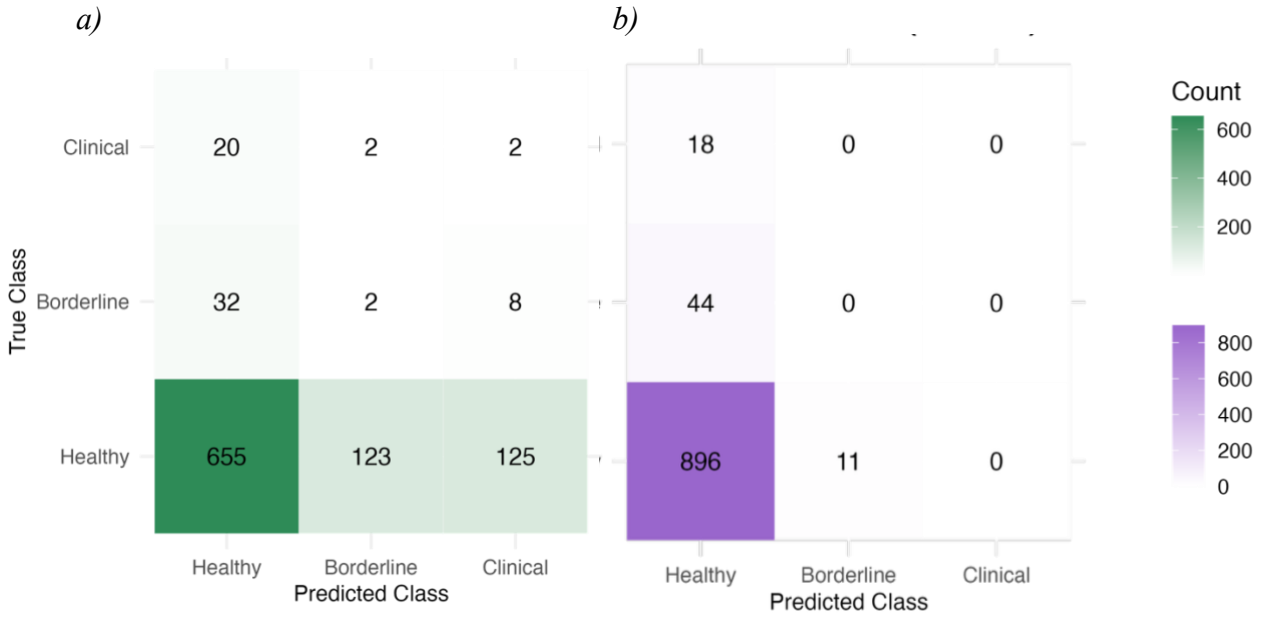
Note: Model performance for each class is shown using predicted probability distributions (top row) and corresponding OvR ROC curves (bottom row). In the one-vs-rest evaluation, the (a) Parent-reported model (CBCL) show predicted probabilities for the healthy class are skewed toward 1; however, the corresponding ROC curve closely follows the diagonal, indicating near-random performance. In contrast, both the borderline and clinical classes have predicted probabilities near zero and ROC curves remains close to chance. For the (b) child-reported model (BPM) the healthy class shows similarly higher predicted probabilities and a ROC curve indicating near-chance. The borderline and clinical class also show low predictied probabilities and low discriminability however, the clinical class, shows some curvature in its ROC curve, suggesting marginally better performance in minority class classification compared to the parent-reported model, although overall class separability remains limited.

These trends are further supported by the confusion matrix (see Figure 11a), which showed that only two cases of borderline and clinical cases were correctly classified for parent-report, with most predictions still concentrated in the healthy category. For the child-report model, there was slight upward curvature in the ROC curve for the clinical group, suggesting some degree of separability. However, the model still failed to predict any true clinical cases, highlighting a disconnect between probabilistic output and actual classification performance (Figure 11b). While the confusion matrix showed more even distribution across predicted labels post-calibration, true positive rates for borderline and clinical remained zero.

Taken together, these findings suggest that sMRI-based features may offer limited utility in accurately distinguishing internalizing symptom severity, particularly when relying on categorical classification frameworks. The slightly improved ROC curvature in the child-report model indicates that children's self-reported symptoms may align somewhat better with underlying neural patterns than parent reports; however, the persistently poor classification performance for minority classes across both models raises questions about the sensitivity of these brain features in capturing clinically meaningful distinctions.

Figure 11

Confusion Matrix for Model Predictions on the Test Set



Note: Confusion matrices display true versus predicted class labels for models after threshold calibration. Color intensity indicates the number of cases per cell. The (a) parent-reported (CBCL) model correctly identified a small number of clinical and borderline cases, but most predictions remained in the healthy category. The (b) child-reported (BPM) model showed more accurate identification of borderline cases but never predicted any cases as clinical. Despite a more balanced distribution, both models remain biased toward predicting the majority (healthy) class.

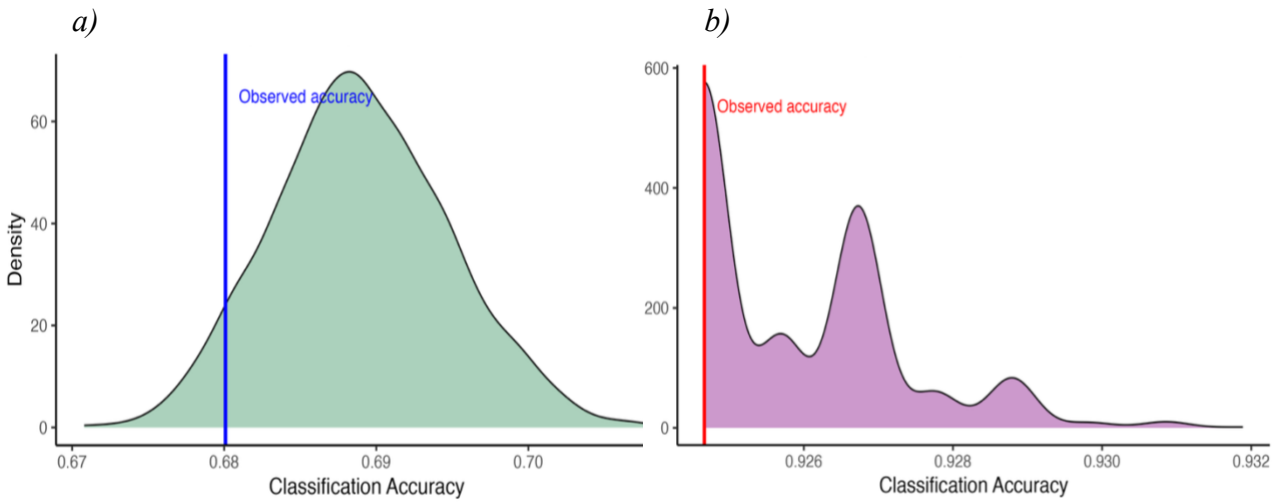
Null Model Validation

To evaluate whether model performance exceeded chance levels, permutation testing was conducted with 1,000 random label shuffles. As shown in Figure 12, the null distribution of accuracy scores is represented by a density curve. The parent-report model (Figure 12a) achieved a high raw classification accuracy of 0.68. However, the observed performance closely overlapped with the null distribution, yielding a permutation-based p-value of 0.953. This suggests that the elevated accuracy was likely driven by class imbalance rather than a meaningful predictive signal. These findings align with previous ROC and confusion matrix analyses, which

indicated that classification success was largely restricted to the majority class. A similar pattern was found for the child-report model (Figure 12b). The model trained on true labels achieved a classification accuracy of 0.93. The null distribution, generated via label permutation, had a mean of 0.55 and a standard deviation of 0.03. The observed accuracy was at the very left of the distribution, yielding a permutation-based p-value of 1.000, indicating that the model's performance was not significantly above chance. The results suggest that the high classification accuracy is likely an artifact of label distribution rather than a reflection of true neural signal.

Figure 12

Permutation Test Comparing Model Accuracy to Null Distribution



Note: Density plots show the null distribution of classification accuracies obtained from 1,000 random permutations of test set labels for the (a) parent-report (CBCL) model and (b) child-report (BPM) model. The observed accuracy using the true (non-permuted) labels is marked by a vertical line (blue for CBCL, red for BPM). For the parent-report model, the observed accuracy fell well within the null distribution, yielding a permutation-based p-value of 0.953. Similarly, the child-report model showed no evidence of predictive performance exceeding chance, with a p-value of 1.000. These results indicate that neither model performed significantly better than random guessing.

Discussion

The primary aim of this study was to evaluate whether structural brain features could be used to predict the severity of OCD-related internalizing symptoms, based on reports from both children and their parents. A secondary aim was to compare the predictive utility of child- versus parent-reported symptoms. To ensure fairness and consistency in this comparison, all models were developed using identical procedures, including class weighting, threshold calibration, and hyperparameter tuning. Initial model performance appeared promising, with high classification accuracy observed across several outcomes. However, further examination suggested that these results were primarily influenced by class imbalance rather than genuine predictive signal. Consequently, the models were limited in their ability to meaningfully address whether sMRI features could predict internalizing symptoms. Similarly, the supplementary question, which anticipated superior performance for predicting child-reported internalizing symptoms, received only limited support. Although the ROC curve for child-reported symptoms exhibited greater curvature, indicating marginally higher discriminative ability and suggesting that children's self-reports may align somewhat more closely with underlying structural neural patterns, this did not result in meaningful improvements in identifying clinically significant cases. These findings suggest that brain features derived from sMRI may not robustly capture the neural correlates of internalizing symptoms, regardless of informant.

This outcome likely reflects several contributing factors. First, the small proportion of participants with elevated symptom severity may have limited the models' ability to detect meaningful associations. Secondly, the use of symptom scores as the target variable may have lacked sufficient granularity to distinguish varying levels of severity. Alternatively, structural brain differences associated with OCD-related internalizing symptoms in children may be too subtle to detect using current methods and features.

The Problem of Imbalanced Data

The results of the XGBoost analyses underscore a fundamental challenge in applying machine learning in low-base-rate mental health contexts: imbalanced class distributions can severely distort model evaluation metrics. Although accuracy is often cited as a primary measure of performance, it can be misleading in contexts where one class (typically the healthy or normative

group) vastly outnumbers others. In such cases, a model may appear highly accurate simply by consistently predicting the majority class, while failing to identify clinically meaningful cases in underrepresented groups. This issue is particularly consequential in mental health research, where accurate detection of borderline and clinical symptom profiles is crucial for screening, diagnosis, and intervention planning.

In theory, XGBoost is well-suited for psychiatric research, where symptom patterns may emerge from multifactorial influences spanning behavioral, biological, and demographic domains. However, in practice machine learning algorithms typically require a sufficient number of examples from each class to effectively learn distinctions. With 194 features and three outcome classes, traditional guidelines recommend at least 10 outcome events per feature to avoid overfitting, although more recent research emphasizes context-specific and simulation-based approaches to sample size planning (Peduzzi et al., 1996; Riley et al., 2019). Applying the 10-events-per-variable rule suggests a minimum of 5,820 observations for balanced class representation ($10 \times 194 \times 3$). While our dataset included 6,460 total observations, only 109 were labeled as clinical cases. Using the same logic, the clinical group alone would require at least 1,940 observations (10×194) to ensure stable performance. Although the sample did not meet this threshold, several best-practice methods, such as class weighting, comprehensive hyperparameter tuning, and threshold calibration were applied to mitigate the effects of class imbalance. Nonetheless, the limited size of the clinical group likely constrained the model's ability to learn robust decision boundaries for that class, contributing to its poor sensitivity.

This limitation stems in part from the nature of the loss function used during model training. XGBoost minimizes log loss (cross-entropy), this function prioritizes the global minimization of prediction error by reducing the average error across all cases (Ng, 2004). Without class rebalancing techniques, they offer minimal learning signal for rare outcomes (He & Garcia, 2009). Consequently, the model learns to be highly confident in classifying healthy cases, while failing to sufficiently learn patterns distinguishing the rarer borderline and clinical groups, this overconfidence for the majority class is clearly visualized in the predicted probabilities of Figure 10. This is likely because the underlying representations learned by the model failed to meaningfully differentiate those groups in the first place.

Challenges of Using Symptom Scores as a Target Variable

Another constraint may be due to the target variable itself, a symptom checklist score. Instruments such as the CBCL and BPM are widely used in child and adolescent mental health research due to their efficiency, standardization, and strong psychometric properties (Achenbach, 2001). However, despite these advantages, such questionnaires may lack the precision needed to capture subtle variations in internalizing symptom severity. This limitation is predominantly salient in non-clinical populations, where symptoms may be subthreshold, situational, or masked by social desirability biases (De Los Reyes & Kazdin, 2005; Youngstrom et al., 2000). Although self-report measures capture the subjective experience of internalizing symptoms, making them particularly valuable for detecting internal distress, they are also vulnerable to underreporting and often show limited agreement with external informants. This can contribute to high intra-individual variability and may reduce the reliability and discriminative power of the symptom data (De Los Reyes et al., 2015).

The findings of Ivankovic et al. (2024) further highlight these concerns. In their study, dimensional ratings of OCD symptoms on the CBCL were compared with clinical OCD diagnoses. While elevated checklist scores were generally associated with diagnosis, they did not reliably differentiate children with clinically significant OCD from those with subclinical symptoms. Importantly, stronger associations were observed for parent-reported obsessions—a domain that is not included in child-report versions of the CBCL and was therefore unavailable in the present study. This distinction underscores a key limitation in our dataset and highlights the broader challenge of informant effects in modeling brain–behavior relationships. While our study aimed to evaluate the predictive utility of both child- and parent-reported symptoms, such asymmetries in questionnaire content may constrain the interpretation of comparative findings.

Limitations of The Model Features

Insufficient Sensitivity of Neuroimaging-Based Input Features

Another limitation of the current study involves the use of tabulated sMRI features as inputs for predictive modeling. T1w sMRI does not directly image brain tissue; rather, it captures radio-frequency signals emitted by hydrogen atoms in water and fat, influenced by their surrounding microenvironment (Weinberger & Radulescu, 2016). Anatomical metrics such as gray matter

volume are derived from intensity contrasts between tissue types and estimated through segmentation algorithms, not by direct visualization of cellular architecture (Ashburner & Friston, 2000). MRI-derived features reflect a composite of neurons, glia, blood vessels, and extracellular components. The resulting signal is modulated by biophysical factors, such as tissue viscosity, perfusion, and magnetic susceptibility, which can be influenced by non-structural variables including hydration status, psychotropic medication, stress, body weight, and substance use (Amianto et al., 2013; Streitbürger et al., 2012; Wang et al., 2022). These factors can introduce signal variability that does not reflect underlying anatomical integrity. In child and adolescent samples, systematic confounders, particularly head motion, may produce apparent differences in brain structure, such as "cortical thinning" or "tissue loss," even when no true pathology exists (Reuter et al., 2015). Furthermore, while sMRI can detect macroscopic structural properties, it cannot resolve the cellular mechanisms, such as synaptic pruning or glial remodeling, that drive these changes. For instance, decreases in grey matter volume are often attributed to synaptic pruning, yet synapses comprise less than 1.5% of cortical volume, and reductions may instead reflect broader processes such as increased intracortical myelination or glial cell loss (Bourgeois & Rakic, 1993; Mills & Tamnes, 2014). Consequently, volumetric sMRI measures may lack the specificity needed to meaningfully link structure to behavior in developing brains.

Moreover, brain development during childhood is nonlinear and regionally asynchronous, meaning that structural differences associated with symptom severity may emerge at different rates depending on the developmental stage of the child (Tamnes et al., 2013). These region-specific developmental trajectories present challenges for sMRI-based ROI analyses, particularly when applying static, adult-derived atlases to child and adolescent samples. For example, cortical grey matter volume typically follows an inverted-U trajectory, peaking in middle childhood and declining through adolescence, and the onset and rate of grey matter changes vary by region (Gilmore et al., 2012; Mills & Tamnes, 2014). Cortical maturation follows a posterior-to-anterior gradient, with earlier development in sensory and parietal regions and later maturation in prefrontal and temporal areas (Tamnes et al., 2013). Furthermore, gyrification decreases from childhood through adolescence due to cortical flattening, reflecting shifts in sulcal depth and width (Alemán-Gómez et al., 2013; Mutlu et al., 2013). Parcellation schemes like the Destrieux atlas, which depend on sulco-gyral anatomy, may therefore misclassify or inconsistently label

cortical regions in children, adding noise to volumetric estimates. Subcortical structures pose similar challenges: despite earlier assumptions of early subcortical maturation, structures such as the amygdala, caudate, and thalamus continue to undergo volumetric change well into adolescence (Herting et al., 2018; Tamnes et al., 2018). These ongoing developmental processes reduce the stability of sMRI-derived subcortical measurements and can lead to misestimation of brain–behavior relationships when relying on single-timepoint, cross-sectional data.

Finally, structural brain differences associated with internalizing symptoms are often subtle and spatially diffuse, making them difficult to detect using coarse-grained sMRI features like global cortical metrics or regional volumes alone (Albaugh et al., 2017). This challenge is further amplified in non-clinical or subclinical populations, where symptom severity is lower, and neural correlates may lie beneath the detection threshold of conventional structural imaging approaches. Recent studies underscore this complexity; for example, Rozovsky et al. (2024) demonstrated that in a transdiagnostic sample of young adults, both increased and decreased cortical thickness in specific regions, such as the left pars opercularis and left inferior temporal gyrus, were differentially associated with depression, anxiety, and mania/hypomania symptom severity. Moreover, these associations were partially mediated by subcomponents of neuroticism, suggesting that personality traits may modulate how structural brain variation relates to internalizing symptoms. These findings highlight that internalizing symptomatology is linked to complex and region-specific patterns of cortical morphology, and underscore the importance of high-resolution, multivariate approaches to capture these nuanced neurobiological correlates in at-risk but non-clinical populations.

Lack of contextual and behavioral data

Although preliminary models included psychosocial and demographic features, the final predictive models focused exclusively on structural brain features. As a result, they did not incorporate contextual or behavioral data that could have enriched the interpretation of internalizing symptom variation. This constitutes another key limitation, given that internalizing symptoms are shaped by ongoing interactions between neurobiological vulnerabilities and environmental exposures, including social stressors, daily routines, and affective response (Insel, 2017). Conventional assessments often fail to capture these influences, relying instead on static, decontextualized symptom ratings obtained in clinical or research settings. This narrow focus

can result in an incomplete picture of behavior and mood, particularly in everyday contexts where symptoms may fluctuate in response to situational demands (Insel, 2017). The absence of real-world behavioral data thus limits the ecological validity of the current models and constrains their capacity to reflect the lived experience of internalizing symptomatology.

Implications and Future Directions

This study underscores several key challenges in applying machine learning to psychiatric prediction tasks, particularly when working with imbalanced class distributions, checklist-based outcome measures, and limited contextual representation. Despite the use of established techniques such as class weighting, hyperparameter tuning, and threshold calibration model sensitivity for clinically significant cases remained poor. These findings suggest that widely used approaches may be inadequate when clinical outcomes are both rare and heterogeneous. Building on the limitations identified in this study, several avenues should be explored to improve the effectiveness of machine learning models in predicting internalizing symptoms in children.

First, addressing the issue of class imbalance would benefit from incorporating sampling techniques such as synthetic oversampling (e.g., SMOTE) and semi-supervised algorithms (Chawla et al., 2002; Zhu & Goldberg, 2009). These methods can help ensure that minority classes are adequately represented during training and can contribute to more equitable and clinically meaningful classification performance. Additionally, increasing the number of clinically affected cases through targeted recruitment or strategic data augmentation may be necessary to meet the data demands of high-dimensional machine learning models (He & Garcia, 2009). Additionally, future work should explore alternative modeling frameworks that better accommodate the challenges of psychiatric prediction.

Second, studies should consider using longitudinal data to differentiate between transient symptom fluctuations and stable clinical trajectories (Dwyer et al., 2018). Modeling symptom change over time may offer a more informative target than single-timepoint scores. Longitudinal designs are also essential for modeling nonlinear, region-specific brain development and distinguishing transient from persistent changes (Tamnes et al., 2018). Given the anatomical variability of the developing brain, children-specific or data-driven parcellation schemes should replace static, adult-derived atlases to reduce misclassification and improve measurement. Where

feasible, incorporating clinician-rated measures or structured diagnostic interviews as ground truth outcomes would further enhance model validity (Achenbach et al., 1987).

Third, expanding the range and quality of input features may significantly improve model performance. Future models should consider multimodal data sources, including neurocognitive assessments, behavioral observations, ecological momentary assessment (EMA; such as in-the-moment mood tracking via smartphones), wearable sensor data (e.g., activity or sleep tracking), and digital phenotyping (e.g., smartphone usage patterns, geolocation and, accelerometer data). These richer data types may capture underlying constructs that are not adequately reflected in standardized rating scales and offer more precise signals for detecting internalizing psychopathology (Bzdok & Meyer-Lindenberg, 2018). The integration of such data would also align with emerging trends in digital mental health, where continuous and passive monitoring can provide real-time insights into symptom dynamics (Insel, 2017).

Importantly, model interpretability must remain a central consideration; clinical decision-making depends not only on accuracy but also on transparency and trust in the model's predictions (Lipton, 2018). Moreover, future research must prioritize real-world validation and reproducibility. Many studies, including the present one, rely on curated or preprocessed datasets that may not fully reflect the variability encountered in applied clinical environments. Validation in external, heterogeneous samples, particularly from clinical or hospital-based populations, is essential for assessing generalizability and translational value (Van Calster et al., 2019). Transparent reporting of model parameters, performance metrics, and code, along with data sharing where possible, is equally important to support reproducibility and cumulative progress in the field (Collins et al., 2015).

Conclusion

This study evaluated the potential of sMRI features to predict OCD-related internalizing symptoms in children, comparing models based on child- and parent-reported data. While classification accuracy was initially high, performance was limited by class imbalance, low symptom prevalence, and the coarse resolution of both imaging and outcome measures. These constraints hindered the detection of clinically meaningful structural neural patterns. These limitations highlight the need for cautious interpretation and suggest that future research may benefit from larger, more balanced samples and the integration of richer, longitudinal, and multimodal approaches.

References

3.3. Tuning the decision threshold for class prediction. (n.d.). Scikit-Learn. Retrieved April 20, 2025, from https://scikit-learn/stable/modules/classification_threshold.html

ABCD Study. (2025, March 14). *MRI Quality Control & Recommended Image Inclusion Criteria*. https://wiki.abcdstudy.org/release-notes/imaging/quality-control.html

Achenbach, McConaughy, S., Ivanova, M., & Rescorla, L. (2017). Manual for the aseba brief problem monitor for ages 6–18 (bpm/6–18). *Burlington: University of Vermont Research Center for Children, Youth, and Families*.

Achenbach, T. M. (2001). *Manual for the ASEBA school-age forms & profiles: Child behavior checklist for ages 6-18, teacher's report form, youth self-report: An integrated system of multi-informant assessment*. ASEBA.

Achenbach, T. M. (2009). *The Achenbach system of empirically based assessment (ASEBA): Development, findings, theory, and applications*. University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M. (2018). Achenbach System of Empirically Based Assessment (ASEBA). In *Encyclopedia of Clinical Neuropsychology* (pp. 26–33). Springer, Cham. https://doi.org/10.1007/978-3-319-57111-9_1529

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/Adolescent Behavioral and Emotional Problems: Implications of Cross-Informant Correlations for Situational Specificity. *Psychol Bull*, *101*(2), 213–232. https://doi.org/10.1037/0033-2909.101.2.213

Achenbach, T. M., McConaughy, S. H., Ivanova, M. Y., & Rescorla, L. A. (n.d.). *Manual for the ASEBA Brief Problem Monitor™ for Ages 6-18 (BPM/6-18)*.

Achenbach, T. M., McConaughy, S. H., Ivanova, M. Y., & Rescorla, L. A. (2011). Manual for the ASEBA brief problem monitor (BPM). *Burlington, VT: ASEBA*, *33*.

Albaugh, M. D., Ducharme, S., Karama, S., Watts, R., Lewis, J. D., Orr, C., Nguyen, T.-V., Mckinstry, R. C., Botteron, K. N., Evans, A. C., & Hudziak, J. J. (2017). Anxious/depressed symptoms are related to microstructural maturation of white matter in typically developing youths. *Dev Psychopathol*, *29*(3), 751–758. https://doi.org/10.1017/S0954579416000444

Alemán-Gómez, Y., Janssen, J., Schnack, H., Balaban, E., Pina-Camacho, L., Alfaro-Almagro, F., Castro-Fornieles, J., Otero, S., Baeza, I., Moreno, D., Bargalló, N., Parellada, M., Arango, C., & Desco, M. (2013). The Human Cerebral Cortex Flattens during Adolescence. *J Neurosci*, *33*(38), 15004–15010. https://doi.org/10.1523/JNEUROSCI.1459-13.2013

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). American Psychiatric Association.

Amianto, F., Caroppo, P., D'Agata, F., Spalatro, A., Lavagnino, L., Caglio, M., Righi, D., Bergui, M., Abbate-Daga, G., Rigardetto, R., Mortara, P., & Fassino, S. (2013). Brain volumetric abnormalities in patients with anorexia and bulimia nervosa: A Voxel-based morphometry study. *Psychiatry Res*, *213*(3), 210–216. https://doi.org/10.1016/j.pscychresns.2013.03.010

Anagnostopoulos, D. C., Korlou, S., Sakellariou, K., Kondyli, V., Sarafidou, J., Tsakanikos, E., Giannakopoulos, G., & Liakopoulou, M. (2016). Comorbid psychopathology and clinical symptomatology in children and adolescents with obsessive-compulsive disorder. *Psychiatriki*, *27*(1), 27.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

ASEBA. (2019, January 14). https://aseba.org/aseba-overview/

Ashburner, J., & Friston, K. J. (2000). Voxel-Based Morphometry—The Methods. *Neuroimage*, *11*(6), 805–821. https://doi.org/10.1006/nimg.2000.0582

Barch, D. M., Albaugh, M. D., Avenevoli, S., Chang, L., Clark, D. B., Glantz, M. D., Hudziak, J. J., Jernigan, T. L., Tapert, S. F., Yurgelun-Todd, D., Alia-Klein, N., Potter, A. S., Paulus, M. P., Prouty, D., Zucker, R. A., & Sher, K. J. (2018). Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description. *Dev Cogn Neurosci*, *32*, 55–66. https://doi.org/10.1016/j.dcn.2017.10.010

Biffen, S. C., Warton, C. M. R., Dodge, N. C., Molteno, C. D., Jacobson, J. L., Jacobson, S. W., & Meintjes, E. M. (2020). Validity of automated FreeSurfer segmentation compared to manual tracing in detecting prenatal alcohol exposure-related subcortical and corpus

callosal alterations in 9- to 11-year-old children. *NeuroImage : Clinical*, *28*, 102368. https://doi.org/10.1016/j.nicl.2020.102368

Bourgeois, J., & Rakic, P. (1993). Changes of synaptic density in the primary visual cortex of the macaque monkey from fetal to adult stage. *J Neurosci*, *13*(7), 2801–2820. https://doi.org/10.1523/jneurosci.13-07-02801.1993

Bragdon, L. B., & Coles, M. E. (2017). Examining Heterogeneity of Obsessive-Compulsive Disorder: Evidence for Subgroups Based on Motivations. *J Anxiety Disord*, *45*, 64–71. https://doi.org/10.1016/j.janxdis.2016.12.002

Breiman, L. (2017). *Classification and Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Brodersen, K. H., Cheng Soon Ong, Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *ICPR*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brownlee, J. (2021, March 21). A Gentle Introduction to XGBoost Loss Functions. *MachineLearningMastery.Com*. https://www.machinelearningmastery.com/xgboost-loss-functions/

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Casey, B. j., Jones, R. M., & Hare, T. A. (2008). The Adolescent Brain. *Annals of the New York Academy of Sciences*, *1124*(1), 111–126. https://doi.org/10.1196/annals.1440.010

Casey, Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … ABCD Imaging Acquisition Workgroup. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *The Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Circulation (New York, N.Y.)*, *131*(2), 211–219. https://doi.org/10.1161/CIRCULATIONAHA.114.014508

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *Neuroimage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dale, A. M., & Sereno, M. I. (1993). Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *J Cogn Neurosci*, *5*(2), 162–176. https://doi.org/10.1162/jocn.1993.5.2.162

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., & Kazdin, A. E. (2005). Informant Discrepancies in the Assessment of Childhood Psychopathology: A Critical Review, Theoretical Framework, and Recommendations for Further Study. *Psychological Bulletin*, *131*(4), 483–509. https://doi.org/10.1037/0033-2909.131.4.483

De Los Reyes, A., & Makol, B. A. (2022). Informant Reports in Clinical Assessment. In *Comprehensive Clinical Psychology* (pp. 105–122). Elsevier. https://doi.org/10.1016/B978-0-12-818697-8.00113-8

de Mathis, M. A., Diniz, J. B., Hounie, A. G., Shavitt, R. G., Fossaluza, V., Ferrão, Y., Leckman, J. F., de Bragança Pereira, C., do Rosario, M. C., & Miguel, E. C. (2013). Trajectory in obsessive-compulsive disorder comorbidities. *Eur Neuropsychopharmacol*, *23*(7), 594–601. https://doi.org/10.1016/j.euroneuro.2012.08.006

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annu Rev Clin Psychol*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Enrico, P., Delvecchio, G., Turtulici, N., Pigoni, A., Villa, F. M., Perlini, C., Rossetti, M. G., Bellani, M., Lasalvia, A., Bonetto, C., Scocco, P., D'Agostino, A., Torresani, S., Imbesi, M., Bellini, F., Veronese, A., Bocchio-Chiavetto, L., Gennarelli, M., Balestrieri, M., … Brambilla, P. (2021). Classification of Psychoses Based on Immunological Features: A Machine Learning Study in a Large Cohort of First-Episode and Chronic Patients. *Schizophrenia Bulletin*, *47*(4), 1141–1155. https://doi.org/10.1093/schbul/sbaa190

Fischl, B., & Dale, A. M. (2000). Measuring the Thickness of the Human Cerebral Cortex from Magnetic Resonance Images. *Proc Natl Acad Sci U S A*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans Med Imaging*, *20*(1), 70–80. https://doi.org/10.1109/42.906426

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Fischl, B., Sereno, M. I., Tootell, R. B. H., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp*, *8*(4), 272–284.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friston, Karl. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Hum. Brain Mapp*, *3*(3), 165–189. https://doi.org/10.1002/hbm.460030303

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R. Z., Heeringa, S., Jernigan, T., Potter, A., Thompson, W., & Zahs, D. (2018). Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience*, *32*, 16–22. https://doi.org/10.1016/j.dcn.2018.04.004

Geller, D. A., M. B. B. S., & March, J., M. D. (2012). Practice Parameter for the Assessment and Treatment of Children and Adolescents With Obsessive-Compulsive Disorder. *J Am Acad Child Adolesc Psychiatry*, *51*(1), 98–113. https://doi.org/10.1016/j.jaac.2011.09.019

Gilmore, J. H., Shi, F., Woolson, S. L., Knickmeyer, R. C., Short, S. J., Lin, W., Zhu, H., Hamer, R. M., Styner, M., & Shen, D. (2012). Longitudinal Development of Cortical and Subcortical Gray Matter from Birth to 2 Years. *Cereb Cortex*, *22*(11), 2478–2485. https://doi.org/10.1093/cercor/bhr327

Good, P. (2000). *Permutation Tests*. Springer. https://doi.org/10.1007/978-1-4757-3235-1

Graybiel, A. M., & Rauch, S. L. (2000). Toward a Neurobiology of Obsessive-Compulsive Disorder. *Neuron*, *28*(2), 343–347. https://doi.org/10.1016/S0896-6273(00)00113-6

Hagler, D. J., Hatton, SeanN., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B. J., Barch, D. M., Harms, M. P., Watts, R., Bjork, J. M., Garavan, H. P., Hilmer, L., Pung, C. J., Sicat, C. S., Kuperman, J., Bartsch, H., Xue, F., … Dale, A. M. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

Haist, F., & Jernigan, T. L. (2023). *Adolescent Brain Cognitive Development Study (ABCD)—Annual Release 5.1*. https://doi.org/10.15154/Z563-ZD24

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Herting, M. M., Johnson, C., Mills, K. L., Vijayakumar, N., Dennison, M., Liu, C., Goddings, A.-L., Dahl, R. E., Sowell, E. R., Whittle, S., Allen, N. B., & Tamnes, C. K. (2018). Development of subcortical volumes across adolescence in males and females: A multisample study of longitudinal changes. *Neuroimage*, *172*, 194–205. https://doi.org/10.1016/j.neuroimage.2018.01.020

Insel, T. R. (2017). Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*, *318*(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Ivankovic, F., Johnson, S., Shen, J., Scharf, J. M., & Mathews, C. A. (2024). Optimization of self- or parent-reported psychiatric phenotypes in longitudinal studies. *Journal of Child Psychology and Psychiatry*. https://doi.org/10.1111/jcpp.14054

Ivarsson, T., Melin, K., & Wallin, L. (2008). Categorical and dimensional aspects of co-morbidity in obsessive-compulsive disorder (OCD). *Eur Child Adolesc Psychiatry*, *17*(1), 20–31. https://doi.org/10.1007/s00787-007-0626-z

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., & Dale, A. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, *30*(2), 436–443. https://doi.org/10.1016/j.neuroimage.2005.09.046

Karcher, N. R., & Barch, D. M. (2021). The ABCD study: Understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*, *46*(1), 131–142. https://doi.org/10.1038/s41386-020-0736-6

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(7), 980–988. https://doi.org/10.1097/00004583-199707000-00021

Kaufman, J., Townsend, L. D., & Kobak, K. (2017). The Computerized Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS): Development and Administration Guidelines. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(10, Supplement), S357. https://doi.org/10.1016/j.jaac.2017.07.770

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, *6*(4), 15:1-15:21. https://doi.org/10.1145/2382577.2382579

KSADS-COMP. (n.d.). Retrieved April 20, 2025, from https://www.ksadslogin.net/ksads-comp/

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Luxburg, U. von, & Schoelkopf, B. (2008). *Statistical Learning Theory: Models, Concepts, and Results* (No. arXiv:0810.4752). arXiv. https://doi.org/10.48550/arXiv.0810.4752

Mills, K. L., & Tamnes, C. K. (2014). Methods and considerations for longitudinal structural brain imaging analysis across development. *Dev Cogn Neurosci*, *9*(C), 172–190. https://doi.org/10.1016/j.dcn.2014.04.004

Mowinckel, A. M., & Vidal-Piñeiro, D. (2020). Visualization of brain statistics with R packages ggseg and ggseg3d. *Advances in Methods and Practices in Psychological Science*, *3*(4), 466–483.

Mutlu, A. K., Schneider, M., Debbané, M., Badoud, D., Eliez, S., & Schaer, M. (2013). Sex differences in thickness, and folding developments throughout the cortex. *Neuroimage*, *82*, 200–207. https://doi.org/10.1016/j.neuroimage.2013.05.076

Nazeer, A., Latif, F., Mondal, A., Azeem, M. W., & Greydanus, D. E. (2020). Obsessive-compulsive disorder in children and adolescents: Epidemiology, diagnosis and management. *Translational Pediatrics*, *9*(Suppl 1), S76.

Nelson, E. C., Hanna, G. L., Hudziak, J. J., Botteron, K. N., Heath, A. C., & Todd, R. D. (2001). Obsessive-Compulsive Scale of the Child Behavior Checklist: Specificity, Sensitivity, and Predictive Power. *Pediatrics*, *108*(1), e14–e14. https://doi.org/10.1542/peds.108.1.e14

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. *Twenty-First International Conference on Machine Learning - ICML '04*, 78. https://doi.org/10.1145/1015330.1015435

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625–632. https://doi.org/10.1145/1102351.1102430

Ojala, M., & Garriga, G. C. (2009). Permutation Tests for Studying Classifier Performance. *2009 Ninth IEEE International Conference on Data Mining*, 908–913. https://doi.org/10.1109/ICDM.2009.108

Pedersen, M. L., Jozefiak, T., Sund, A. M., Holen, S., Neumer, S.-P., Martinsen, K. D., Rasmussen, L. M. P., Patras, J., & Lydersen, S. (2021). Psychometric properties of the Brief Problem Monitor (BPM) in children with internalizing symptoms: Examining baseline data from a national randomized controlled intervention study. *BMC Psychology*, *9*(1), 185. https://doi.org/10.1186/s40359-021-00689-1

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, *49*(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Prechelt, L. (1998). *Early Stopping—But When?* (G. B. Orr & K.-R. Müller, Eds.; Vol. 1524, pp. 55–69). Germany: Springer Berlin / Heidelberg. https://doi.org/10.1007/3-540-49430-8_3

Ren, H., Wang, X., Wang, S., & Zhang, Z. (2019). Predict Fluid Intelligence of Adolescent Using Ensemble Learning. In K. M. Pohl, W. K. Thompson, E. Adeli, & M. G. Linguraru (Eds.), *Adolescent Brain Cognitive Development Neurocognitive Prediction* (pp. 66–73). Springer International Publishing.

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage*, *107*, 107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006

Rifkin, R., & Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, *5*(Jan), 101–141.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*, *38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Rozovsky, R., Bertocci, M., Iyengar, S., Stiffler, R. S., Bebko, G., Skeba, A. S., Brady, T., Aslam, H., & Phillips, M. L. (2024). Identifying tripartite relationship among cortical thickness, neuroticism, and mood and anxiety disorders. *Scientific Reports*, *14*(1), 8449. https://doi.org/10.1038/s41598-024-59108-1

Ryan, M. (2025). *Machine Learning for Tabular Data.* (1st ed.). Manning Publications Co. LLC.

Saad, L. O., do Rosario, M. C., Cesar, R. C., Batistuzzo, M. C., Hoexter, M. Q., Manfro, G. G., Shavitt, R. G., Leckman, J. F., Miguel, E. C., & Alvarenga, P. G. (2017). The Child Behavior Checklist—Obsessive-Compulsive Subscale Detects Severe Psychopathology and Behavioral Problems Among School-Aged Children. *Journal of Child and Adolescent Psychopharmacology*, *27*(4), 342–348. https://doi.org/10.1089/cap.2016.0125

Saragosa-Harris, N. M., Chaku, N., MacSweeney, N., Guazzelli Williamson, V., Scheuplein, M., Feola, B., Cardenas-Iniguez, C., Demir-Lira, E., McNeilly, E. A., Huffman, L. G., Whitmore, L., Michalska, K. J., Damme, K. S., Rakesh, D., & Mills, K. L. (2022). A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. *Developmental Cognitive Neuroscience*, *55*, 101115. https://doi.org/10.1016/j.dcn.2022.101115

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., Gur, R. C., & Gur, R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *Neuroimage*, *60*(1), 623–632. https://doi.org/10.1016/j.neuroimage.2011.12.063

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press. https://doi.org/10.7551/mitpress/8291.001.0001

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *Neuroimage*, *22*(3), 1060–1075. https://doi.org/10.1016/s1053-8119(04)00188-0

Segonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically Accurate Topology-Correction of Cortical Surfaces Using Nonseparating Loops. *IEEE Trans Med Imaging*, *26*(4), 518–529. https://doi.org/10.1109/TMI.2006.887364

Shephard, E., Stern, E. R., van den Heuvel, O. A., Costa, D. L. C., Batistuzzo, M. C., Godoy, P. B. G., Lopes, A. C., Brunoni, A. R., Hoexter, M. Q., Shavitt, R. G., Reddy, Y. C. J., Lochner, C., Stein, D. J., Simpson, H. B., & Miguel, E. C. (2021). Toward a neurocircuit-based taxonomy to guide treatment of obsessive–compulsive disorder. *Mol Psychiatry*, *26*(9), 4583–4604. https://doi.org/10.1038/s41380-020-01007-8

Shmueli, G. (2011, January 5). *To Explain or to Predict?* arXiv.Org. https://doi.org/10.1214/10-STS330

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging*, *17*(1), 87–97. https://doi.org/10.1109/42.668698

Storch, E. A., Kay, B., Wu, M. S., Nadeau, J. M., & Riemann, B. (2017). Suicidal and Death Ideation among Adults with Obsessive-Compulsive Disorder Presenting for Intensive Intervention. *Ann Clin Psychiatry*, *29*(1), 46–53. https://doi.org/10.1177/104012371702900109

Streitbürger, D.-P., Möller, H. E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M. L., & Mueller, K. (2012). Investigating Structural Brain Changes of Dehydration Using Voxel-Based Morphometry. *PLoS One*, *7*(8), e44195–e44195. https://doi.org/10.1371/journal.pone.0044195

Tamnes, C. K., Bos, M. G. N., van de Kamp, F. C., Peters, S., & Crone, E. A. (2018). Longitudinal development of hippocampal subregions from childhood to adulthood. *Dev Cogn Neurosci*, *30*, 212–222. https://doi.org/10.1016/j.dcn.2018.03.009

Tamnes, C. K., Walhovd, K. B., Dale, A. M., Østby, Y., Grydeland, H., Richardson, G., Westlye, L. T., Roddey, J. C., Hagler, D. J., Due-Tønnessen, P., Holland, D., & Fjell, A. M. (2013). Brain development and aging: Overlapping and unique patterns of change. *Neuroimage*, *68*, 63–74. https://doi.org/10.1016/j.neuroimage.2012.11.039

Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, *14*(3), 659–665. https://doi.org/10.1109/TKDE.2002.1000348

Tisdall, M. D., Hess, A. T., Reuter, M., Meintjes, E. M., Fischl, B., & van der Kouwe, A. J. W. (2012). Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magnetic Resonance Medicine*, *68*(2), 389–399. https://doi.org/10.1002/mrm.23228

Tisdall, M. D., Reuter, M., Qureshi, A., Buckner, R. L., Fischl, B., & van der Kouwe, A. J. W. (2016). Prospective motion correction with volumetric navigators (vNavs) reduces the bias and variance in brain morphometry induced by subject motion. *Neuroimage*, *127*, 11–22. https://doi.org/10.1016/j.neuroimage.2015.11.054

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., McLernon, D. J., Moons, K. G. M., Steyerberg, E. W.,

Van Calster, B., van Smeden, M., Vickers, A. J., & On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

van den Heuvel, O. A., Boedhoe, P. S. W., Bertolin, S., Bruin, W. B., Francks, C., Ivanov, I., Jahanshad, N., Kong, X.-Z., Kwon, J. S., O'Neill, J., Paus, T., Patel, Y., Piras, F., Schmaal, L., Soriano-Mas, C., Spalletta, G., van Wingen, G. A., Yun, J.-Y., Vriend, C., … ENIGMA-OCD working group. (2022). An overview of the first 5 years of the ENIGMA obsessive-compulsive disorder working group: The power of worldwide collaboration. *Human Brain Mapping*, *43*(1), 23–36. https://doi.org/10.1002/hbm.24972

van den Heuvel, O. A., van Wingen, G., Soriano-Mas, C., Alonso, P., Chamberlain, S. R., Nakamae, T., Denys, D., Goudriaan, A. E., & Veltman, D. J. (2016). Brain circuitry of compulsivity. *European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology*, *26*(5), 810–827. https://doi.org/10.1016/j.euroneuro.2015.12.005

Wald, L., Schmitt, F., & Dale, A. (2001). Systematic spatial distortion in MRI due to gradient non-linearities. *NeuroImage (Orlando, Fla.)*, *13*(6), 50–50. https://doi.org/10.1016/S1053-8119(01)91393-X

Wang, Z., Fontaine, M., Cyr, M., Rynn, M. A., Simpson, H. B., Marsh, R., & Pagliaccio, D. (2022). Subcortical shape in pediatric and adult obsessive-compulsive disorder. *Depression and Anxiety*, *39*(6), 504–514. https://doi.org/10.1002/da.23261

Weinberger, D. R., & Radulescu, E. (2016). Finding the Elusive Psychiatric "Lesion" With 21st-Century Neuroanatomy: A Note of Caution. *Am J Psychiatry*, *173*(1), 27–33. https://doi.org/10.1176/appi.ajp.2015.15060753

Weisz, J. R., Doss, A. J., & Hawley, K. M. (2005). Youth Psychotherapy Outcome Research: A Review and Critique of the Evidence Base. *Annual Review of Psychology*, *56*(Volume 56, 2005), 337–363. https://doi.org/10.1146/annurev.psych.55.090902.141449

Wells, W. M., Viola, P., Atsumi, H., Nakajima, S., & Kikinis, R. (1996). Multi-modal volume registration by maximization of mutual information. *Med Image Anal*, *1*(1), 35–51. https://doi.org/10.1016/S1361-8415(01)80004-9

White, N., Roddey, C., Shankaranarayanan, A., Han, E., Rettmann, D., Santos, J., Kuperman, J., & Dale, A. (2010). PROMO: Real-time prospective motion correction in MRI using image-based tracking. *Magn. Reson. Med*, *63*(1), 91–105. https://doi.org/10.1002/mrm.22176

Wu, M. S., Geller, D. A., Schneider, S. C., Small, B. J., Murphy, T. K., Wilhelm, S., & Storch, E. A. (2019). Comorbid Psychopathology and the Clinical Profile of Family Accommodation in Pediatric OCD. *Child Psychiatry Hum Dev*, *50*(5), 717–726. https://doi.org/10.1007/s10578-019-00876-7

XGBoost. (2022). https://xgboost.readthedocs.io/en/stable/parameter.html

XGBoost Developers. (2022). *XGBoost Parameters—Xgboost 3.1.0-dev documentation*. https://xgboost.readthedocs.io/en/latest/parameter.html#general-parameters

xgboost Grid Search—R. (n.d.). Retrieved April 20, 2025, from https://kaggle.com/code/silverstone1903/xgboost-grid-search-r

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*(6), 1038–1050. https://doi.org/10.1037/0022-006X.68.6.1038

Zhu, X., & Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. *3*(1), 1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006

Zucker, R. A., Gonzalez, R., Feldstein Ewing, S. W., Paulus, M. P., Arroyo, J., Fuligni, A., Morris, A. S., Sanchez, M., & Wills, T. (2018). Assessment of culture and environment in the Adolescent Brain and Cognitive Development Study: Rationale, description of measures, and early data. *Developmental Cognitive Neuroscience*, *32*, 107–120. https://doi.org/10.1016/j.dcn.2018.03.004

Appendix I

Data Splitting and Class Distributions

To ensure appropriate model training, calibration, and evaluation, the dataset was randomly divided into training (70%), calibration (15%), and test (15%) subsets using stratified random sampling via the create DataPartition function from the caret package in R. A fixed seed (1) was used to ensure reproducibility. Stratification was based on internalizing class labels (healthy, borderline, clinical) to maintain class proportions across splits.

Table 1.

Class Distribution Across Dataset Splits for Internalizing Domain

CBCL					
Split	Healthy	Borderline	Clinical	Total	Minority %
Full	6126	222	112	6460	5.2%
Training	4301	152	69	4522	4.9%
Calibration	922	28	19	969	4.9%
Test	903	42	24	969	6.8%
BPM					
Split	Healthy	Borderline	Clinical	Total	Minority %
Full	6043	308	109	6460	~2.5%
Training	4231	217	74	4522	~2.7%
Calibration	905	47	17	969	~2.8%
Test	907	44	18	969	~2.3%

Note. "Minority %" indicates the proportion of samples in the borderline and clinical categories combined.

Table 2.

Class Distribution Across Dataset Splits for Externalizing Domain

CBCL					
Split	Healthy	Borderline	Clinical	Total	Minority %
Full	6356	66	38	6460	~1.6%
Training	4455	44	23	4522	~1.5%
Calibration	953	12	4	969	~1.7%
Test	948	10	11	969	~2.2%
BPM					
Split	Healthy	Borderline	Clinical	Total	Minority %
Full	6096	144	21	6261	~2.5%
Training	4266	102	15	4383	~2.7%
Calibration	913	24	2	939	~2.8%
Test	917	18	4	939	~2.3%

Note. Children missing externalizing T-scores due to incomplete item responses.

Appendix II

Summary of Linear Regression Models Predicting Internalizing Symptoms

Linear regression models were conducted to assess the contribution of demographic, psychosocial, and neuroimaging features for predicting internalizing symptoms as reported by parents and children. The outcome variables were T-scores from the CBCL for parent reports and the BPM for child self-reports. Model fit statistics for each regression are summarized below.

Table 1

Summary of Linear Regression Models Predicting Parent Reported Internalizing Symptoms

Model	Residual SE	R^2	Adjusted R^2	F (df)	p-value
Age, Sex, Race/Ethnicity	9.70	0.016	0.014	12.07 (6, 4515)	< .001
Latent SES, Social Risk, Perinatal Risk	9.73	0.010	0.009	14.56 (3, 4518)	< .001
sMRI	9.68	0.060	0.019	1.45 (191, 4330)	< .001
All Features	9.62	0.073	0.031	1.73 (196, 4325)	< .001

Note. Residual SE = Residual Standard Error. R^2 = Coefficient of Determination. All models used CBCL Internalizing T-score as the outcome variable. Four linear regression models were conducted, each model includes a different combination of demographic, latent psychosocial, and neuroimaging variables. Key model fit indices are reported, including residual standard error, R^2, adjusted R^2, F-statistics, and associated p-values.

Table 2

Summary of Linear Regression Models Predicting Child Reported Internalizing Symptoms

Model	Residual SE	R^2	Adjusted R^2	F (df)	p-value
Age, Sex, Race/Ethnicity	5.11	0.005	0.004	4.13 (6, 4515)	<.001
Latent SES, Social Risk, Perinatal Risk	5.08	0.017	0.016	25.63 (3, 4518)	<.001
sMRI	5.07	0.060	0.018	1.44 (191, 4330)	<.001
All Features	5.03	0.074	0.032	1.77 (196, 4325)	<.001

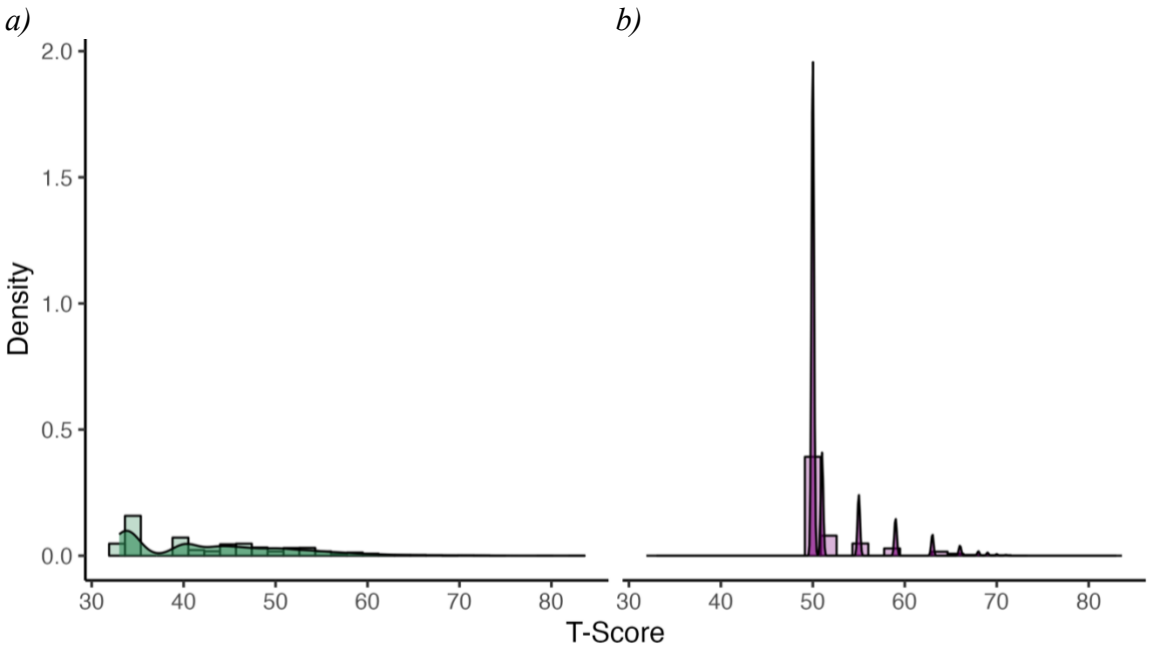
Appendix III

XGBoost Models Predicting Externalizing Symptoms

The externalizing symptoms target variable exhibited more severe class imbalance issues than the internalizing symptoms, with a predominance of participants classified as "normal," and comparatively fewer in "borderline" and "clinical" categories. These results are provided here for transparency.

Figure 1

Distribution of Externalizing Outcomes



Note: Externalizing T-scores, which closely resembles the distribution of internalizing scores.

Table A

Best Tuning Parameters for Externalizing Symptoms – XGBoost Models

Parameter	CBCL	BPM
Booster	Gbtree	Gbtree
Objective	Multi:softprob	Multi:softprob
Evaluation Metric	mlogloss	mlogloss
Max Depth	3	3
Min Child Weight	5	1
Eta	0.1	0.05
Gamma	0.5	0.5
Subsample	0.6	0.9
Colsample by Tree	0.9	0.9
Number of Classes	3	3
Best number of rounds	150	50
Best multiclass Log Loss	1.091149	1.08137

Note: These hyperparameters were obtained through cross-validation to minimize multiclass log loss.

Table B

Model Performance Metrics for Externalizing Symptoms

	CBCL				BPM			
Model Configuration	Default	With Weights	Tuned + Argmax	Tuned + Threshold-Cal.	Default	With Weights	Tuned + Argmax	Tuned + Threshold-Cal.
Overall Performance								
Accuracy	0.3478	0.3498	0.9835	0.9546	0.3323	0.3365	0.9723	0.886
95% CI	0.3178–0.3787	0.3198–0.3808	0.9733–0.9905	0.9395–0.9668	0.3022–0.3634	0.3063–0.3678	0.9597–0.9818	0.864–0.9057
Cohen's Kappa	0.0113	0.0084	0	-0.0173	-0.0106	0.005	0	-0.0328
Class Balance								
Balanced Accuracy	0.5524	0.5545	0.5	0.4879	0.4248	0.5038	0.5	0.462
Sensitivity (Borderline)	0.4	0.4	0.0	0.0	0.1667	0.3333	0.0	0.0
Sensitivity (Clinical)	0.7273	0.5455	0.0	0.0	0.5	0.5	0.0	0.0
Specificity (Borderline)	0.6705	0.6621	1.0	0.9875	0.683	0.6743	1.0	0.924
Specificity (Clinical)	0.6754	0.6858	1.0	0.9885	0.6545	0.6599	1.0	0.984
Class-Level Discrimination								
Precision (Borderline)	0.0125	0.0122	NaN	0.0	0.0102	0.0196	NaN	0.0
Precision (Clinical)	0.0251	0.0195	NaN	0.0	0.0062	0.0063	NaN	0.0
F1 Score (Borderline)	0.0236	0.0236	0.0	0.0	0.0	0.0	0.0	0.0
F1 Score (Clinical)	0.0248	0.0248	0.0	0.0	0.0	0.0	0.0	0.0

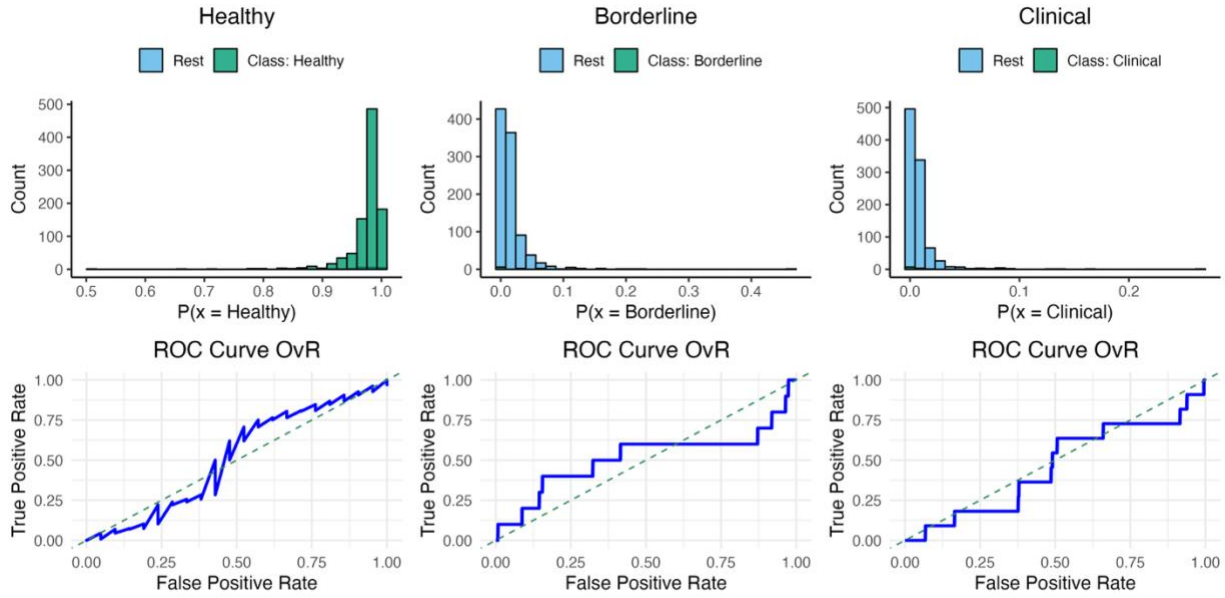
Note: The calibrated threshold approach improved overall accuracy substantially. However, class-level discrimination for minority classes (borderline and clinical) remained poor. The

extreme imbalance has overwhelmed the model's capacity to generalize effectively across classes, highlighting the difficulty in modeling low-prevalence categories even with calibrated thresholds.

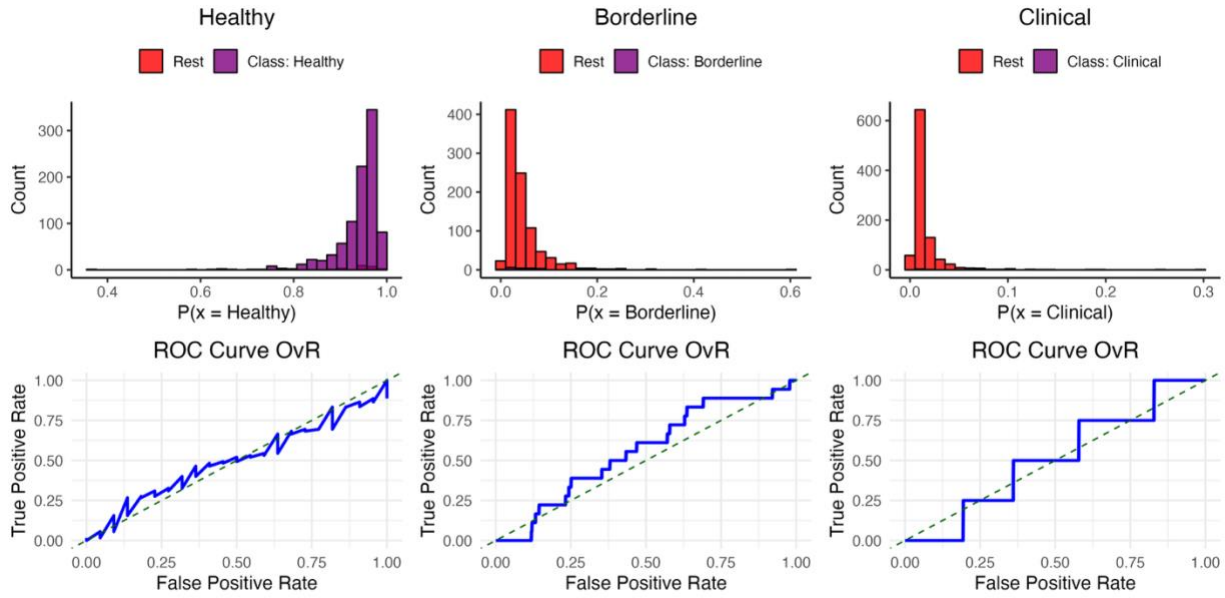
Figure 2

ROC Curves and Class Probability Distributions for Externalizing Models

a)



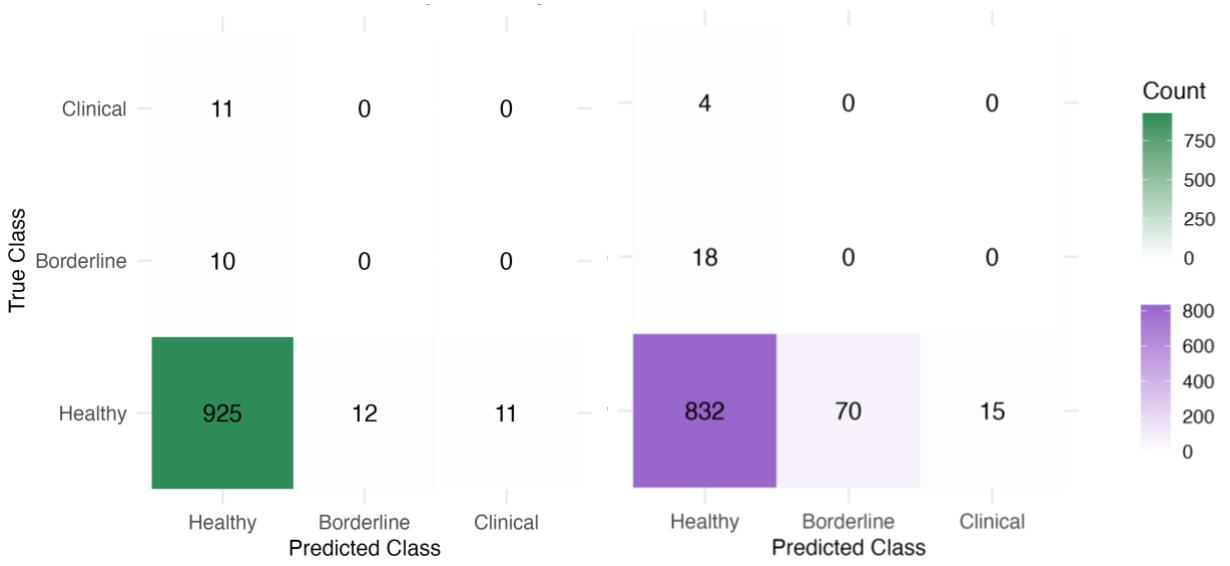
b)



Note: Class probability distributions and ROC curves for externalizing symptom classifications across three severity levels (healthy, borderline, clinical), using OvR approach. The models are trained separately on (a) parent-report (CBCL) and (b) child-report (BPM) data. In both models, the probability distributions for the healthy class showed a strong peak near 1, indicating high model confidence. In contrast, predictions for the borderline and clinical classes were heavily skewed toward lower probabilities (near 0), indicating a lack of model confidence. The corresponding ROC curves further reflect this pattern: both models achieved the highest area under the curve (AUC) for the healthy class, although performance for all categories were near chance levels. Overall, the models struggled to differentiate between borderline and clinical levels of externalizing symptoms, likely due to class imbalance, overlapping feature distributions, and limited sensitivity of structural brain features alone.

Figure 3

Confusion Matrix for Model Predictions for The Externalizing Models



Note: Confusion matrices for the test set classification performance of the externalizing symptom models, displaying results from the (a) parent-report (CBCL) model in green and the (b) child-report model in purple. Both matrices reveal a strong bias toward predicting the healthy class, with limited correct identification of borderline and clinical cases. In the parent-report model, nearly all healthy cases were correctly classified, while only 11 clinical and 10 borderline cases were present in the test set, and none were correctly identified. Instead, all clinical and borderline

cases were misclassified as healthy, reflecting a complete failure to detect higher symptom severity categories. The child-report model exhibited a slightly higher rate of misclassification within the healthy group. Although this suggests a marginally more distributed prediction pattern, the model still failed to correctly classify any minority cases. Overall, both models demonstrated poor sensitivity to clinically significant externalizing symptoms.