Discussion

The primary aim of this study was to evaluate whether structural brain features could be used to predict the severity of OCD internalizing symptoms as reported by both youths and their parents. In addition to answering the primary research question, a secondary research question involved comparing whether structural brain features were more predictive of internalizing symptoms when reported by youths as compared to their parents. To support this comparative analysis, all models were trained and evaluated using the same procedures, including class weighting, threshold calibration, and hyperparameter tuning, to ensure fairness and consistency across informants and symptom domains. Results show that while initial performance metrics suggested high classification accuracy, further analysis revealed that this performance was largely driven by class imbalance, rather than genuine predictive power. As such, the models were ultimately deemed unfit to answer our primary and secondary research questions.

Main Findings

The results of the multiclass classification 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.

The primary hypothesis that sMRI features could meaningfully predict internalizing symptoms was not supported when evaluated using stringent performance metrics. Similarly, the secondary hypothesis, which anticipated superior performance from self-report models, received only limited support. Although prediction model of child reported internalizing symptoms demonstrated modest improvements in ROC curve patterns, these did not translate into meaningful gains in identifying clinically relevant cases.

Several factors likely contributed to the models’ poor performance. First, the class imbalance severely constrained the models’ ability to learn from minority cases, even with class weighting and calibration, the models continued to default to the healthy category. Second, the output variable, limited to symptom scores, may not have contained enough information to meaningfully distinguish between symptom severity levels. Third, the models lacked contextual or behavioral data that might better capture risk patterns, such as information on social environment, functioning, or life stressors. Taken together, these results echo longstanding concerns about the low signal-to-noise ratio in community-based samples and the difficulty of extracting clinically actionable insights from neuroimaging features alone. These challenges are compounded by well-documented issues in the literature, including the poor generalizability of structural MRI classifiers and the complexities introduced by informant discrepancies in youth psychopathology assessment (Arbabshirani et al., 2017).

The Problem of Imbalanced Data

Despite the theoretical strengths of algorithms like XGBoost in modeling complex relationships, their performance is constrained when trained on datasets that lack sufficient representation of clinically significant but rare cases. In theory, XGBoost is well-suited to psychiatric research, where symptom patterns may emerge from multifactorial influences spanning behavioral, biological, and demographic domains. However, in practice, the algorithm's performance is constrained by the structure and distribution of the training data. These limitations are not surprising, given that 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 predictor variable 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 post hoc suggests a minimum of 5,820 observations for balanced class representation (10 × 194 × 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 fell short of this threshold, we applied several best-practice methods—such as class weighting, threshold calibration, and rigorous hyperparameter tuning—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, standard loss functions like cross-entropy prioritize the global minimization of prediction error, it aims to reduce 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 is likely because the underlying representations learned by the model failed to meaningfully differentiate those groups in the first place. The slightly better sensitivity observed in the CBCL-based model may reflect greater parent-report consistency, but overall classification performance remained far from clinically actionable.

Challenges of Using Symptom Scores for Prediction

Another limitation may stem from the nature of the target variable itself, symptom checklist scores, while standardized and widely used, may lack the granularity needed to distinguish between subtle variations in internalizing symptom severity. Instruments such as the CBCL and BPM are popular in youth mental health research due to their efficiency and strong psychometric properties (Achenbach, 2001). However, they often fail to capture the full complexity of symptom presentation, particularly in non-clinical populations where expressions of internalizing distress may be subthreshold, situational, or masked by social desirability biases (De Los Reyes & Kazdin, 2005; Youngstrom et al., 2000). These limitations are especially pronounced in self-report measures, which are prone to underreporting and inconsistent agreement with external informants (De Los Reyes et al., 2015). These factors contribute to high intra-individual variability and frequent underreporting of internalizing symptoms, which in turn diminishes the reliability and discriminative power of the symptom data (Achenbach et al., 1987).

Recent findings by Ivankovic et al. (2024) support this concern. In their study, the authors compared dimensional ratings of OCD symptoms on the CBCL with clinical OCD diagnoses (Ivankovic et al., 2024). They found that while elevated scores on symptom checklists were generally associated with an OCD diagnosis, they did not reliably distinguish youth with clinically significant OCD from those with milder or subclinical symptom patterns. Notably, parent-reported obsessions showed a stronger relationship with diagnosis, suggesting that internalizing symptoms may need to reach a threshold of external observability to reflect clinically meaningful distress. These findings reinforce the rationale for examining internalizing symptoms in predictive models but also highlight the importance of considering informant effects, particularly when modeling brain–behavior relationships.

Lack of contextual and behavioral data

Third, the models did not incorporate contextual or behavioral data that could have provided a richer understanding of the individual’s psychosocial environment. This is a notable limitation, as internalizing symptoms are shaped by complex interactions among personal, social, and environmental influences (Insel, 2017).

Implications for Future Research

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, threshold calibration, and hyperparameter tuning, 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.

Future Directions

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 youth. These directions pertain to both the methodological refinement of modeling approaches and the enhancement of the underlying data used to train and evaluate such models.

First, addressing the issue of class imbalance remains a critical priority. In many psychiatric datasets, clinically significant cases are underrepresented. Future research 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. While ensemble-based methods like XGBoost offer strong baseline performance, deep learning approaches or hybrid models that combine interpretability and flexibility may provide added value, particularly when working with complex, high-dimensional datasets (Shwartz-Ziv & Armon, 2022).

Second, improvements to the outcome variable are necessitated. Future 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. Furthermore, the integration of multiple informants, rather than analyzing youth- and parent-reported symptoms in isolation, could help capture a more comprehensive view of symptomatology and reduce the noise introduced by informant discrepancies (De Los Reyes & Makol, 2022). 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), wearable sensor data, and digital phenotyping. 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). Finally, 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, McLernon, van Smeden, Wynants, & Steyerberg, 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).

-also youth brain might be very noisy because of developmental changes

Conclusion

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., 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

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

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

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

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

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

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

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

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

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Twenty-First International Conference on Machine Learning - ICML ’04, 78. https://doi.org/10.1145/1015330.1015435

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

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

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Med, 17(1), 230–230. https://doi.org/10.1186/s12916-019-1466-7

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