SSM - Population Health 32 (2025) 101886

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

What machine learning teaches us about depression prediction across the life course: An exploratory comparison of predictive models

Rafael Geurgas [a,*], Saul J. Newman [b,c,d], Evelina T. Akimova [a,e,f], Katherine N. Thompson [a], Robbee Wedow [a,e,f,g]

[a] Department of Sociology, Purdue University, USA
[b] Centre for Longitudinal Studies, University College, UK
[c] University College, University of Oxford, UK
[d] Institute of Population Aging, University of Oxford, UK
[e] Department of Statistics, Purdue University, USA
[f] Center on Aging and the Life Course (CALC), Purdue University, USA
[g] Department of Medical and Molecular Genetics, Indiana University School of Medicine, USA

ARTICLE INFO

Keywords:
Depression
Prediction
Machine learning
Polygenic scores

ABSTRACT

Identifying individuals at risk for depression early is important for preventing long-term mental health issues. However, the variability in depression severity, duration, and triggers complicates predictions. This study explores whether machine learning models can outperform traditional methods, like Logistic Regression, in predicting self-reported depressive symptoms and clinical depression during adolescence and adulthood. We applied five machine learning models with varying complexity levels – Logistic Regression, Decision Tree, XGBoost, Support Vector Machine, and Neural Networks – using data from a nationally representative longitudinal study of the U.S., which tracked participants for 20 years. The models were trained with early-life predictors (ages 12–18) from Wave I, including environmental factors (family, school, health) and genetic predispositions (polygenic scores) from Wave IV. Models were evaluated on their ability to predict depressive symptoms and clinical diagnoses in both adolescence and adulthood. After evaluating the performance of all five models, XGBoost emerged as the most effective, with a 0.02 increase in ROC-AUC compared to the benchmark Logistic Regression model. While this is a slight performance improvement, overall, Logistic Regression performs about as well as many of our ML models. Early-life data showed strong predictive value for depressive symptoms and clinical diagnoses in adolescence and adulthood, highlighting adolescence as a critical period. Polygenic scores do not add predictive power when combined with environmental data. Feature importance analyses identified self-perception and physical health as key predictors of depressive symptoms, while trauma and life-changing events were more influential for clinical depression.

1. Introduction

Mental health is currently a global challenge and a public health concern. Mental health problems are among the leading causes of overall morbidity (Vos et al., 2015) and disability (Lozano et al., 2012), and are further linked to suicide risk (Appleby et al., 2017).

Depression is one of the most common and widely studied mental health conditions. Depression has been described as the *"common cold"* of mental health (Gitterman, 1991), a metaphor that underscores its widespread prevalence and impact. It affects over 300 million people worldwide with detrimental consequences for individuals' health, relationships, and life outcomes (Kessing et al., 2023). This burden is especially pronounced in adolescents and young adults, whereby depression and anxiety are leading contributors to overall sickness and disability (World Health Organization, 2023).

Despite growing attention and public health efforts, depression often goes undiagnosed until it becomes severe. Conventional approaches to identifying at-risk groups, such as screenings based on a limited set of risk factors (e.g., family history, trauma exposure, or socioeconomic disadvantage), have had only moderate success in predicting future depressive episodes. The complexity of causes playing a role in the development of depression poses a core challenge to the task of tackling its different stages. Depression has both internal and external risk factors that represent a complex net of interconnected causes arising from the interplay of genetic predispositions, psychological, social, and

* Corresponding author.
E-mail address: rgeurgas@purdue.edu (R. Geurgas).

https://doi.org/10.1016/j.ssmph.2025.101886
Received 7 August 2025; Received in revised form 17 November 2025; Accepted 18 November 2025
Available online 19 November 2025

other environmental influences (Fried & Nesse, 2017; Matheson et al., 2006; McPherson & Armstrong, 2006; Monroe & Simons, 2005; Reading & Reynolds, 2001). For example, two individuals with similar symptom levels may have very different underlying risk factor profiles. This complexity makes it challenging for traditional statistical methods and clinical judgment alone to predict who will develop depression accurately. This requires an interdisciplinary approach with an integrative, predictive focus that can handle high-dimensional data and capture subtle risk patterns before the onset of disorder.

Importantly, adolescence is a critical developmental window as approximately half of all mental health disorders in adulthood have their onset by age 18 (Kessler et al., 2005). Yet many adolescent depression cases remain undetected or untreated, allowing symptoms to carry into later life. Adolescence is characterized by rapid biological, psychological, and social changes, and exposure to new stressors; these can heighten vulnerability to depression during this formative period (Mills et al., 2014). Failure to address mental health needs in youth has cascading effects, impairing educational achievement, social relationships, and health into adulthood (Costello et al., 2003; Patel et al., 2007). This underscores the urgency of improving early identification of depression risk. The ability to predict an emerging depressive episode is imperative for developing timely and effective interventions that mitigate the severity of symptoms and improve the overall quality of life of individuals at risk (Hirschfeld, 2012).

Recent advances in data science have opened new avenues for improving depression prediction. In particular, machine learning (ML) methods hold promise for analyzing large datasets with numerous predictors, uncovering nonlinear relationships and interactions that conventional regression models may overlook. Likewise, recent advances in genomics have introduced a unique opportunity for the use of polygenic scores (PGS), which are aggregate indices of genetic liability for a trait derived from genome-wide association studies (GWAS), offering a novel source of predictive information alongside established psychosocial risk factors. These innovations raise an exciting possibility: combining rich longitudinal data on adolescents' environments and behaviors with PGS indicators and applying state-of-the-art ML algorithms could substantially improve our ability to predict depression in adolescence and adulthood. At the same time, the use of ML and genetic data presents new challenges. Models must be rigorously evaluated for their accuracy and generalizability in a diverse population, and careful attention is needed to interpretability issues (e.g., can we understand the predictions and underlying risk drivers).

This paper is situated at the intersection of these developments, aiming to advance an interdisciplinary understanding of depression development. By integrating computational methods, genetic data, and social-environmental measures, our study asks: *Can combining genetic and environmental predictors in an ML framework improve the prediction of depression from adolescence and adulthood?* To answer this question, we employ four ML methods: Decision Tree (Breiman et al., 2017; Quinlan, 1986), eXtreme Gradient Boosting — XGBoost (Chen & Guestrin, 2016), Support Vector Machine — SVM (Cortes & Vapnik, 1995), and one deep learning method (Neural Networks (LeCun et al., 2015; Rumelhart et al., 1986)), which we compare to the standard benchmark Logistic Regression approach (Cox, 1958), most used in social sciences. We utilize data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative longitudinal survey of the U.S. population (Harris et al., 2019), to systematically evaluate these multiple approaches, integrating roughly 100 genetic and non-genetic predictors. We also pay particular attention to how predictive performance can vary over developmental time, differ by sex, or depend on how depression is defined and measured. Additionally, we explore the key factors that contribute most significantly to predicting depression. Our approach extends prior research by integrating genomic data and assessing the viability of applying a machine learning framework within a social science context.

By bridging computational, statistical genomics, and social science approaches, our study demonstrates how ML can complement (rather than replace) conventional analytical frameworks, yielding more nuanced insights into determinants of depression. Through this interdisciplinary approach, we aim to clarify the potential and limitations of advanced predictive models in mental health and to inform strategies for the early identification of depression risk that bridge biological and social contexts.

2. Background

2.1. Framing depression: From public health burden to interdisciplinary inquiry

Depression is a prevalent, complex, multifaceted mental health disorder that demands interdisciplinary attention. While its significance as a public health priority is widely acknowledged, there is less consensus on how best to understand its causes, measurement, and progression over time. Traditional efforts to define and address depression have spanned clinical, psychological, and epidemiological domains, yet growing awareness of its social and biological underpinnings has broadened the scope of inquiry. As research moves beyond within-disciplinary lenses, new frameworks are emerging to capture the layered, dynamic nature of depression, ranging from neurobiology to family relationships, and from molecular genomics to structural inequality. Here, we motivate our inquiry by tracing key disciplinary contributions to the study of depression, highlighting the ongoing shift toward integrative approaches to model large-scale data.

Beyond its sheer prevalence, depression is characterized by considerable complexity in both its presentation and measurement. Major depressive disorder (MDD) is defined by a constellation of symptoms. It includes persistent low mood, loss of interest or pleasure (anhedonia), and disturbances in sleep, appetite, or concentration that last over a sustained period (American Psychiatric Association, 2013). However, individuals meeting the diagnostic criteria for MDD can exhibit widely divergent symptom patterns. The standard criteria allow for numerous combinations of symptoms, and large patient studies have documented over a thousand unique symptom profiles among those diagnosed (Fried & Nesse, 2015). Such heterogeneity complicates the assessment and classification of depression. Common measurement instruments, from clinician-administered interviews to self-report scales, provide standardized severity scores but may mask important qualitative differences between individuals (Beck et al., 1961; Kroenke et al., 2001). Moreover, depression frequently co-occurs with other mental health conditions, especially anxiety disorders, blurring diagnostic boundaries and further complicating assessment (Kessler et al., 2003). Consequently, scholars continue to debate the conceptualization of depression; whether it represents a singular construct or an umbrella term for a spectrum of related conditions and how best to capture its nuances in research (Horwitz, 2007; Ruggero et al., 2019).

Accordingly, multiple disciplinary perspectives have been applied to understand depression's etiology and persistence. Seeing depression through the lenses of the psychoanalytic model, our starting point becomes aspects of its development. For example, we would focus on how and what early childhood circumstances made us vulnerable to depression. On this assumption of the developmental nature of depression, more effective interventions could also be built. Psychoanalytic therapy aims to reveal insights from unpleasant experiences, as another assumption of this theory holds that uncovered insights are a source of healing power (Peterson, 2009). The family systems model likewise highlights the importance of childhood in relation to depression. However, this notion is based on the assumption that mental health problems are driven by disturbances in the family (Jacobson & Addis, 1993). Both of these theories provide the fundamentals of modern talk therapies, such as psychodynamic and interpersonal therapy (IPT), which are effective in treating (de Mello et al., 2005; Shedler, 2010). If we stand

for cognitive-behavioral theory and its understanding of depression, we would emphasize the role of stressful situations along with the ways of thinking as driving forces. We would encourage therapies that teach people to develop adaptive habits and greater control over thinking processes. This approach provides the foundation for the big family of modern Cognitive-behavioral therapy (CBT) techniques (Feldman, 2007). One of the most important contributions of these psychological approaches is insights into the intraindividual mechanisms that contribute to the development of depression. Such insights continue to play a critical role in the development of individual therapies, further contributing to the clinical utility of psychotherapies (Nathan, 2007).

One of the most important shifts in the understanding of depression is rooted in a biological revolution in psychiatry that occurred during the last quarter of the twentieth century following an empirical success to demonstrate the role of biology in depression development (Schwartz & Corcoran, 2009). Then, biological and biomedical research has probed the genetic and neurophysiological underpinnings of depression. At the molecular level, GWAS have begun identifying numerous genetic variants linked to depression, each conferring a small increase in risk and collectively underscoring the polygenic architecture of the disorder (Wray et al., 2018). Neurobiological investigations have implicated dysfunction in neurotransmitter systems (such as serotonin and norepinephrine), dysregulation of the hypothalamic–pituitary–adrenal (HPA) axis that governs stress hormone release, and other physiological perturbations. For example, many depressed individuals exhibit elevated cortisol levels, a biological signature of chronic stress activation (Stetler & Miller, 2011), and some research implicates heightened inflammatory immune responses in the development or maintenance of depressive symptoms (Slavich & Irwin, 2014). These findings support a view of depression as an illness rooted in brain–body interactions, wherein neurochemical imbalances, endocrine stress responses, and immune system changes contribute to the emergence of pathological mood states. The biological perspective has not only informed pharmacological treatments (such as antidepressants targeting specific neurotransmitter pathways) but also continues to explore biomarkers that might improve diagnosis or predict treatment response.

Complementing the psychological and biological domains, sociological and demographic perspectives situate depression within a broader socio-environmental context. Research in social epidemiology and medical sociology has documented how social determinants, including socioeconomic disadvantage, gender inequality, racial/ethnic stressors, and exposure to trauma or violence, can affect mental health. Depression can be triggered by factors such as educational attainment (Lee, 2011), marital status (Kessler & Essex, 1982; Pearlin & Johnson, 1977), and economic recessions (Frasquilho et al., 2015; Jahoda, 1988). Brown and Harris (2012) demonstrated that acute life events (such as the death of a loved one or loss of a job), in combination with a lack of social support, can trigger depressive episodes, especially among individuals already facing social adversity. Subsequent studies have consistently found higher rates of depression among those experiencing poverty, unemployment, social isolation, or discrimination, underscoring the influence of social stress and inequality in the etiology of depression (Ezzy, 1993; Fryers et al., 2003). Such findings reinforce that depression is not only a biomedical condition but also a socially situated phenomenon, one deeply intertwined with community and structural conditions.

There has been a growing movement toward interdisciplinary frameworks integrating insights across the psychological, biological, and social domains. An important example is the biopsychosocial model, which advocates for a holistic approach that considers biological processes, individual psychology, and socio-environmental context in tandem (Engel, 1977). In other words, the biopsychosocial perspective emphasizes that depressive disorders emerge from the dynamic interplay of neurobiological factors, psychological characteristics, and social context. Therefore, we grounded our study in this framework, incorporating biological factors (genetic risk), psychological and behavioral factors (e.g., mental health history, lifestyle, stress), and social factors (family and school) into a unified predictive modeling approach. This kind of interdisciplinary strategy aligns with broader calls in health research to bridge the gap between *causes of the causes* (social-environmental determinants) and underlying biology (Mabry et al., 2008). We further employ ML techniques not only for their predictive power but also as a methodological bridge to operationalize the holistic nature of the biopsychosocial model.

2.2. Machine learning and its role in depression research

ML has rapidly emerged as a powerful asset in depression research, offering novel insights at both the population and individual clinical levels. By leveraging large, complex datasets, ML models can detect subtle patterns and interactions that are difficult to discern with traditional analytical methods. This capacity allows researchers to uncover new facets of depression etiology and prognosis without being limited to a priori hypotheses (Ahmed et al., 2019; Shatte et al., 2019). In effect, ML-based techniques extend and complement established research methods, enabling analyses ranging from macro-scale public health trends to micro-scale individual predictions. The value of these approaches lies in their flexibility and scalability: ML can integrate diverse types of data – ranging from social media text and wearable sensor readings to electronic health records (EHRs) and genomic profiles – and identify complex associations within them, thereby providing a more nuanced understanding of depression's determinants and trajectories. Importantly, these advances do not replace the insights gained from conventional methods; rather, they augment and enrich them by revealing patterns that might otherwise remain hidden.

Recent advances in machine learning have not only improved predictive performance but also enhanced our ability to interpret complex models. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) allow researchers to "peek inside the black box" and gain insights into how features contribute to model predictions. These interpretability methods enable the bridging of the gap between predictive accuracy and explanatory power, which is particularly valuable in social science research. For instance, Lundberg and Lee (2017) introduced SHAP values as a unified framework for interpreting model predictions, demonstrating their effectiveness in explaining output from a wide range of machine learning models, including tree-based methods and deep learning. Similarly, Ribeiro et al. (2016) introduced LIME, which approximates complex models locally with simpler, interpretable ones, further enhancing the transparency of machine learning algorithms.

One of the most extensively studied contexts is social media, where user-generated content serves as a rich source of behavioral and linguistic signals. In a recent scoping review, 54 studies were identified that used social media platforms to detect depression and anxiety using natural language processing (NLP) and predictive modeling techniques (Shatte et al., 2019). Chaudhury et al. (2013) demonstrated that support vector machines (SVMs) trained on Twitter data labeled with CES-D scores could detect depression-indicative posts with precision above 80%, catalyzing a wave of research in this domain.

Beyond social media, ML models have been applied to data from wearable sensors, voice and speech patterns, mobile phone usage, and digital phenotyping to infer mood states in real-time (Low et al., 2020; Mohr et al., 2017). In clinical research, multimodal ML approaches incorporating cortisol levels, brain activity (EEG), and structural neuroimaging have been used to distinguish depression from anxiety, achieving classification accuracies ranging from 67% to over 90% depending on the input modality and model used (Frick et al., 2014; Gao et al., 2020). Automated speech analysis, for instance, has shown promise in identifying psychiatric symptoms, including depressive affect and cognitive slowing, across clinical and non-clinical samples (Low et al., 2020).

EHRs represent another area for ML-driven depression prediction, enabling early diagnosis and risk stratification based on medical history, prescriptions, and comorbidities (Del Pozo-Banos et al., 2021). Similarly, recent efforts have explored the utility of omics data – including genomics and proteomics – in identifying biological signatures of depression using supervised ML models (Hirschfeld et al., 2022). While these approaches demonstrate high predictive potential, they also face challenges related to data availability, generalizability, and clinical implementation.

From an algorithmic perspective, studies commonly report performance metrics exceeding 70% accuracy for classification tasks involving depression and anxiety. While deep learning approaches such as convolutional neural networks (CNNs) have achieved impressive results (e.g., up to 96% classification accuracy in some studies Ahmed et al., 2019), traditional models like random forests, SVMs, and decision trees remain widely used due to their interpretability and robustness (Chekroud et al., 2016; Chung et al., 2020; Sau & Bhakta, 2019). The optimal algorithm often depends on the input features and problem structure, with heterogeneous findings across data types and study populations.

These advances underscore the versatility of ML in modeling depression across diverse mental health outcomes and contexts, including schizophrenia, bipolar disorder, and PTSD (Shatte et al., 2019). At the same time, they emphasize the need for ongoing innovation in real-world mental health prediction, particularly in general population samples where access to clinical data may be limited. In contrast to the growing body of work utilizing online behavior or clinical biomarkers, our study employs ML on a large, nationally representative survey dataset, providing new insights into individual-level predictors of adolescent depression.

2.3. Integrating polygenic scores in depression prediction

Advances in genomics over the past decade have enabled researchers to quantify genetic susceptibility to depression through PGS. A PGS aggregates the small contributions of thousands of genetic variants across the genome, each weighted by effect sizes identified in GWAS. Depression, like most common mental disorders, has a highly polygenic architecture: no single gene has a strong effect, but many genes each contribute a tiny increase in risk. Family and twin studies estimate that roughly 30%–40% of the liability for major depressive disorder is genetic in origin (Jansson et al., 2004; Keyes, 2005; McGue & Christensen, 2003; Sullivan et al., 2000). Yet because the genetic influence is spread over multiple loci, early attempts to use genetic information for prediction were limited. Initial PGS for depression explained less than 1% of variance in depressive symptoms (Musliner et al., 2015).

As GWAS sample sizes have grown (now comprising hundreds of thousands of individuals), the predictive power of depression PGS has improved, but it remains modest. Contemporary PGS for major depression can explain on the order of only 2%–3% of the variance in depression liability (Plomin & Von Stumm, 2022). This is notably lower than PGS for some other traits (for instance, schizophrenia PGS is around 6% or educational attainment — around 11% of variance) (Plomin & Von Stumm, 2022), owing to the still elusive and complex genetics of depression.

Given this modest standalone predictive power, a critical question is whether incorporating a PGS for depression into multifactorial risk models yields any meaningful improvement in prediction above and beyond traditional risk factors. There is growing evidence that PGS adds incremental value when combined with clinical or environmental variables, albeit typically a small increment. For example, a recent large study in the UK Biobank examined risk models for depression that combined a PGS with early-life environmental risk factors (Lu et al., 2023). The model integrating genetic and environmental predictors achieved an Area Under the Receiver Operating Characteristic Curve (AUC) of about 0.68 for predicting major depression, compared to 0.66 with environment alone and 0.63 with PGS alone (Lu et al., 2023). In other words, knowing someone's PGS in addition to their childhood exposures slightly improved the ability to identify who would develop depression. While this improvement was statistically significant, the authors noted that the magnitude of the gain was relatively small (on the order of 1–3 percentage points in AUC). This pattern – small additive benefit of PGS – appears to be a common finding in emerging studies. It suggests that genetic data can sharpen risk stratification slightly, perhaps helping to flag some high-risk individuals who would not be identified by environmental risk factors alone. However, the bulk of explainable risk still comes from non-genetic factors, at least with current PGS accuracy.

3. The present study: An interdisciplinary approach

By combining theoretical grounds, analytic, and empirical tools from sociology, psychology, computational genomics, and computer science, we aim to provide insights that none of these fields could achieve in isolation. Interdisciplinary and multilevel approaches are critical for understanding and improving population health (Mabry et al., 2008). Our work embodies this ethos by vertically integrating data across multiple levels, utilizing information from DNA to individual behaviors and social contexts, and applying novel analytical techniques to predict depression.

In particular, we use a rich nationally-representative Add Health study to predict early adulthood depression based on adolescent-level predictors. By leveraging this longitudinal cohort, we can examine the prediction of depression in a population-based setting (as opposed to a clinical sample), thereby enhancing the relevance of our findings to public health. The longitudinal design further allows us to test predictive models at different life stages – for instance, predicting depressive outcomes in adolescence versus young adulthood – to see how prediction accuracy and important predictors might change as individuals age. However, we pay critical attention to the adolescent period because it is a developmental window where the first onset of depression often occurs (Backes & Bonnie, 2019). Moreover, supervised ML models allow us to integrate a wide range of survey-based and biological data, including polygenic predictors.

A central aim of our study is to compare multiple ML models for predicting depression, to determine whether more complex models yield performance gains over simpler approaches in this context. We focus on five modeling approaches that span a range from traditional, interpretable methods in statistics to advanced, flexible state-of-the-art deep learning algorithms: (1) Logistic Regression — a baseline statistical model often used in epidemiology for binary outcomes, included for its interpretability and as a benchmark; (2) Decision Tree — a tree-based model that makes decisions by recursively splitting data into subsets based on feature values to predict outcomes or classify data.; (3) Support Vector Machine (SVM) — a kernel-based classifier known for finding optimal boundaries in high-dimensional feature space; (4) Extreme Gradient Boosting (XGBoost) — a powerful boosting algorithm that builds trees sequentially to minimize errors, often leading to high accuracy in structured data; and (5) Neural Network (Multilayer Perceptron) — a feed-forward artificial neural net that can learn complex function mappings given enough data. We directly assess the trade-off between interpretability and predictive performance by training and evaluating these models on the same prediction tasks. Our systematic comparison addresses the inconsistencies in past studies, which have used different data or outcome definitions. All models are applied to the Add Health data under identical conditions. In doing so, we aim to build models that are not only predictive but also broadly generalizable across different sex population subgroups. Our emphasis on interpretability, subgroup performance, and public relevance helps advance the field toward practical applications in adolescent mental health screening and early intervention.

Our analytical framework follows the following stages with respect to the environmental measures included in addition to genetic predictions: First, using existing knowledge on the determinants of depression collected from the literature (Remes et al., 2021; Solmi et al., 2023; Thompson et al., 2024), we build a ML model that takes into account the complex interaction between predictors while avoiding the challenges of endogeneity (Verhagen, 2023). Previous studies have often relied on inferential statistical methods to test single association theories driven by theories about specific determinants of depression (Bzdok et al., 2018; Yarkoni & Westfall, 2017). However, the call for a comprehensive assessment of depression risk and the application of statistical learning algorithms for prediction has been widely echoed, also as a promising tool to aid clinical practice (Iniesta et al., 2016; Iyortsuun et al., 2023; Richter et al., 2021; Yarkoni & Westfall, 2017).

Second, we examined 97 predictor variables to maximize the prediction of symptoms of depression. This allowed us to analyze patterns and how they interact. Integrating ML and deep learning techniques into predicting depression represents a promising opportunity to process large amounts of high-dimensional data, where the number of predictors exceeds the number of observations (Breiman, 2001). Using powerful deep learning algorithms, such as neural networks, can excel in identifying patterns and interactions within large datasets that are often imperceptible using traditional statistical methods (Jordan & Mitchell, 2015). Including a large number of environmental and genetic variables to improve predictive care can help better personalize risk assessments for depression (Richter et al., 2021).

Third, we compare the performance of the ML models to determine which one is most suitable for the task, based on the highest AUC. Additionally, we examine which features (questions) are most influential in the prediction.

Fourth, we leverage both symptomatic indicators and self-reported clinical diagnoses of depression available in the Add Health dataset to address the limitations associated with measurement error discussed earlier. By utilizing these complementary sources of information, we aim to capture a more comprehensive and robust representation of depressive symptomatology, thus mitigating concerns related to single-source bias or underreporting in self-assessments alone. Subsequently, we compare the symptomatic predictions with the self-reported clinical diagnosis results.

Finally, and importantly, we integrate the PGS for depression into the best model to test its added predictive value in this social-environmental context. In practice, we compare model performance across three scenarios: with the inclusion of PGS alongside other traditional predictors (e.g., demographics, socioeconomic status, life events, etc.), without PGS, and using only PGS data without any environmental variables. This enables us to quantify the incremental contribution of genetic information and assess whether this contribution differs by algorithm or sex. For instance, a complex nonlinear model might capture gene–environment interactions in a way that makes the PGS more useful, or conversely, a logistic regression might already capture what the PGS offers through correlated environmental measures. By systematically evaluating this across algorithms, we provide new evidence on how genetics might be integrated with ML for mental health. If the PGS adds little to no improvement, that is a valuable finding indicating that current genetic data might best be left out of predictive tools (or that more powerful PGS are needed). If it does add, we can determine under what conditions and by how much, guiding future interdisciplinary work.

This study takes an initial step in assessing the viability of machine learning models for predicting depression, offering a new perspective beyond traditional approaches. Demonstrating this potential is a necessary foundation for future work that could lead to clinically useful tools to support earlier detection and timely intervention.

4. Data

We used data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative U.S. cohort followed over five waves spanning two decades (Harris et al., 2019). Wave I (1994–1995) surveyed adolescents aged 12–18 in grades 7–12, with follow-ups in Wave II (one year later), Wave III (ages 18–26), Wave IV (24–32), and Wave V (32–42;2016–2018) (Dennis et al., 2022).

Add Health provides rich longitudinal data on demographic, familial, social, school, behavioral, psychosocial, cognitive, and health-related factors. Data were collected through participant and parent surveys, as well as contextual information from schools, neighborhoods, and residential areas. In-home assessments included physical measures, medication use, biomarkers, and genetic data. Detailed study design information is available in Harris et al. (2019).

This study draws on data from two waves of Add Health: the Wave I In-Home Interview (N_{W1} = 20,745), which includes comprehensive environmental and psychosocial data collected during adolescence, and the Wave IV In-Home Interview (N_{W4} = 15,701), which includes self-reported depressive symptoms and PGS for a subsample of respondents. Because PGS are derived from GWAS conducted predominantly in European populations, and predictive accuracy can be compromised in diverse samples due to population stratification and differences in linkage disequilibrium patterns, we restricted genetic analyses to participants of European ancestry. This subsample comprised 5731 individuals (N_{PGS}), or 62.8% of the genotyped respondents. Analytic sample sizes varied depending on the wave and type of data used: analyses drawing on adolescent predictors used the full Wave I sample, while those incorporating depressive symptoms or genetic data used the Wave IV and European ancestry subsample, respectively. The smallest overlapping group, which combined data across waves or sources, determined the analytic sample.

5. Measures

5.1. Self-report depressive symptoms

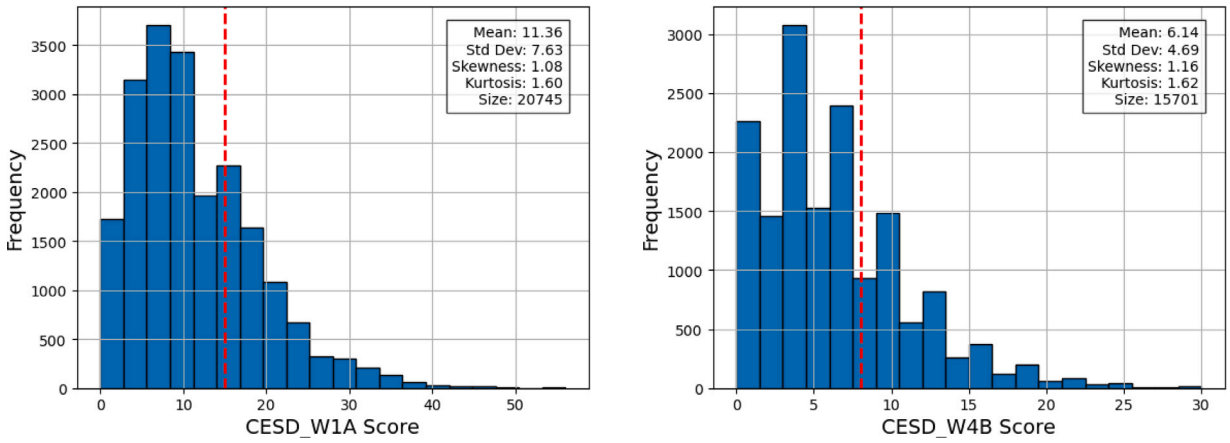
In Waves I and IV, participants completed items from the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977). Wave I included 19 items during adolescence, while Wave IV used a 10-item version in adulthood. From these sources, we derived two depressive symptom indicators: CES-D-based symptoms in adolescence (Wave I) and CES-D-based symptoms in adulthood (Wave IV).

CES-D items measured the frequency of depressive symptoms over the past week using a four-point scale (0–3), with higher scores indicating more severe symptoms. Total scores ranged from 0 to 57 in Wave I and 0 to 30 in Wave IV. Participants with up to four missing items were retained, and scores were rescaled based on the number of completed items to ensure comparability (Radloff, 1977).

The distribution of the CES-D is shown in Fig. 1. Panel (a) for Wave I (19 items) and Panel (b) for Wave IV (10 items). Adjusted thresholds of 15 (Wave I) and 8 (Wave IV) indicate elevated depressive symptoms, based on the original CES-D guidelines.

To construct outcome variables, we categorized participants into binary groups reflecting depressive symptom severity. For both Wave I and Wave IV, individuals with CES-D scores below the wave-specific threshold were classified as having lower depressive symptoms. In comparison, those with scores at or above the threshold were classified as having higher depressive symptoms.

In Wave I, analysis of the 19-item CES-D scale showed that 75% of participants were classified as having lower depressive symptoms, while 25% met the threshold for higher depressive symptoms. In Wave IV, using the 10-item CES-D scale, 21% of participants were classified as having higher depressive symptoms, and 79% fell below the threshold.



(a) The 19 CES-D score from Wave I.

(b) The 10 CES-D score from Wave IV.

Fig. 1. The depression score for Add Health Wave I and Wave IV participants. The vertical dashed line is the recommended threshold for higher and lower depressive symptoms.

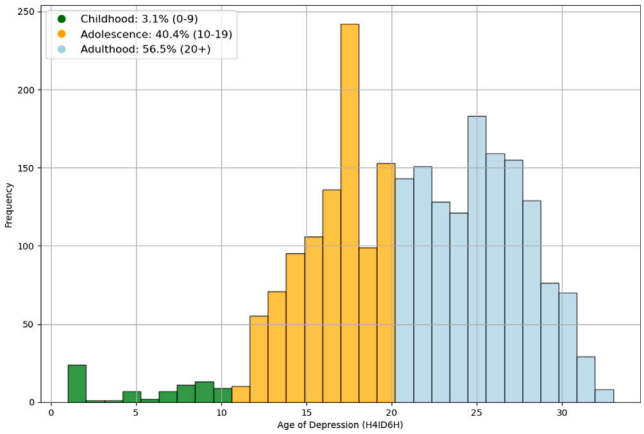


Fig. 2. Age of Depression. Calculated from the retrospective report.

5.2. Moving beyond symptomatic data

In Wave IV, participants were asked: "At what age did a doctor, nurse, or other healthcare provider diagnose you with depression?" (question H4ID6H). A total of 2394 (15%) individuals reported having received a clinical diagnosis of depression at some point in their lives.

As shown in Fig. 2, 3.1% of the total diagnoses occurred during childhood, 40.4% during adolescence, and 56.5% during adulthood. These findings suggest that while the majority of depression diagnoses emerged in adulthood, a substantial proportion was already evident by adolescence, highlighting the importance of early identification and intervention.

To evaluate how early-life psychosocial and environmental factors predict clinically diagnosed depression, we applied the best-performing algorithm – identified during the depressive symptom analysis – to the full Wave IV sample, which includes both sexes. We focused on two distinct outcomes based on retrospective self-reported clinical diagnosis data from the Add Health study: (1) depression diagnosed during adolescence, and (2) depression diagnosed in adulthood.

5.3. Environmental factors

Add Health Wave I includes a wide range of environmental variables, distributed across 39 separate questionnaires covering diverse domains of adolescent life. To guide our selection of predictors, we followed frameworks based on the social determinants of mental health, as well as the biological, psychological, and social determinants of depression (Kirkbride et al., 2024; Remes et al., 2021; Thompson et al., 2024), ensuring comparability and grounding our analysis in established empirical practices. Depressive symptoms were not included in the prediction, mental health-related items were incorporated, and their impact was subsequently evaluated in sensitivity analyses.

The complete list of environmental factors used in the analysis is provided in the Supplementary material (Section 1.1). In total, we included 72 environmental variables for male participants and 75 for female participants. The difference reflects the presence of sex-specific questions in the survey instruments, including items related to menstrual health, experiences of sexual coercion, and pregnancy outcomes.

5.4. Polygenic scores

We included 19 PGS for male participants and 21 for female participants. The difference reflects the inclusion of two sex-specific scores in females: age at menarche and age at menopause. It is worth noting that PGS for MDD was also included as a predictor. We included a broad range of polygenic scores that have been phenotypically associated with depression. A complete list of all polygenic scores used is provided in the Supplementary material (Section 1.2).

We used the standardized polygenic scores available in Add Health (Release 2). These scores were constructed by the Add Health research team using published GWAS summary statistics and were standardized within genetic ancestry groups to account for population stratification. All scores were used as provided, without re-estimation or parameter modification, following the procedures described in the Add Health Polygenic Scores User Guide.[1]

6. Methodology

To assess the effectiveness and applicability of ML models, we developed five supervised learning models to predict depression outcomes, ranging from traditional statistical models to state-of-the-art deep learning methods. These included four machine learning algorithms – logistic regression, decision trees, eXtreme Gradient

[1] https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/WaveIVPGSRelease2UserGuide.pdf

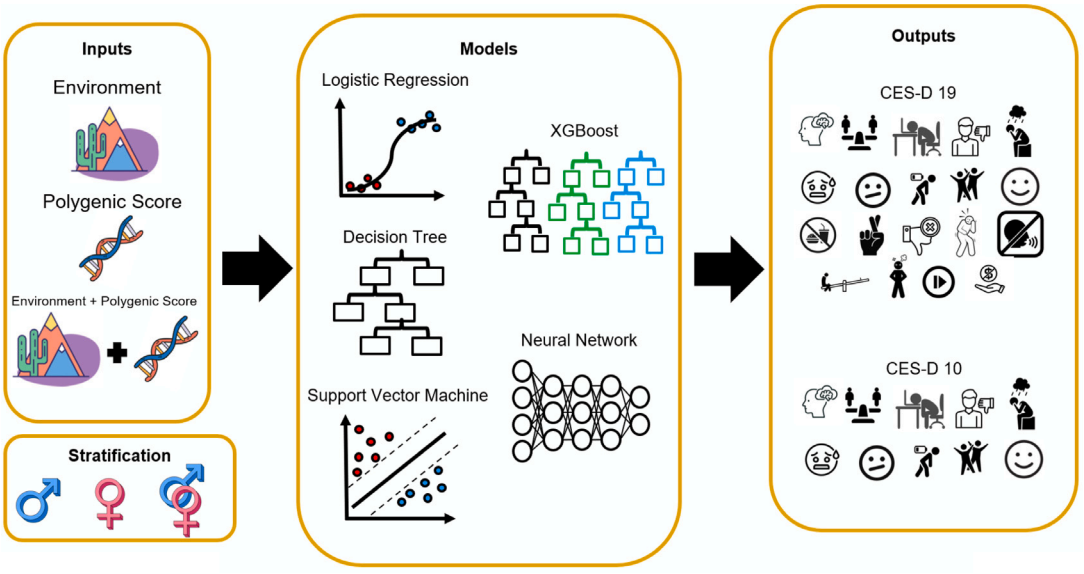


Fig. 3. Illustration of the models used. Three inputs, three sex stratifications, six ML models (SVM with two kernels), and two outputs. The best performance model was extended to compare with clinical depression and PGS.

Boosting (XGBoost), and support vector machines (SVMs) – and one deep learning model, a feedforward neural network. For more details on the models and the study pipeline, refer to the Supplementary material, Section 2. Each model was trained to predict two distinct binary outcomes: higher versus lower depressive symptoms, using CES-D-19 scores from Wave I and CES-D-10 scores from Wave IV.

All models were trained using data from Wave I. Thus, analyses involving Wave I data are considered concurrent (Wave I → Wave I), while those involving Wave IV data are considered prospective (Wave I → Wave IV). Participants not present in Wave IV were excluded only from the prospective analysis.

Model performance was evaluated using two standard metrics: the Area Under the Receiver Operating Characteristic Curve and classification accuracy. The receiver operating characteristic (ROC) curve is a widely used tool for evaluating the performance of binary classifiers. It plots the true positive rate (TPR) against the false positive rate (FPR) across all possible decision thresholds. The area under the ROC curve (AUC) provides a single scalar summary of classifier performance, representing the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. An AUC of 0.5 indicates performance equivalent to random chance, whereas higher values reflect greater discriminative ability. These metrics enabled us to compare model performance across different algorithmic approaches, feature sets, and sex-specific subsamples.

The selection of models is based on their capacity to capture both linear and non-linear relationships within the data. Each model offers distinct advantages. Logistic Regression is favored for its simplicity, interpretability, and low computational cost (Boateng et al., 2019). In contrast, decision tree-based and gradient boosting models are well-suited for modeling complex, non-linear patterns in the input features (Bader et al., 2024; Xin & Ren, 2022). Support Vector Machines (SVMs), employing both linear and Radial Basis Function (RBF) kernels, are particularly effective in handling high-dimensional datasets (Son et al., 2010; Unnikrishnan et al., 2016). Neural networks, in theory, possess a greater potential for predictive power due to their ability to model complex, hierarchical data relationships. This makes them particularly promising for tasks such as depression prediction. However, a significant drawback is that neural networks might underperform compared to simpler models when the dataset is limited or lacks sufficient diversity, as they require large volumes of data to effectively capture intricate patterns.

This factorial design resulted in 54 total models (6 algorithms × 3 input sets × 3 sex strata), as shown in Fig. 3, SVM is tested twice, once with a linear kernel and once with a non-linear kernel. To further explore the best-performing model, we conducted four additional analyses. First, we evaluated its performance against two clinical depression outcomes: self-reported diagnoses in adolescence and adulthood. Those values are reconstructed from self-reported diagnoses in Wave IV. Second, we examined feature importance to identify the most influential risk factors (concurrent and prospective). Third, we assessed the contribution of genetic information by comparing model performance with and without polygenic scores. Fourth, we conducted a sensitivity analysis to assess how the results were affected by removing (a) physical and mental health items closely related to depression and (b) only including the polygenic score for major depressive disorder.

The code is available at the author's GitHub.[2]

6.1. Pre-processing

The following pre-processing steps were performed on all subsets N_{W1}, N_{W4}, and N_{PRS}.

The first step is to impute the missing data. Any missing data (NaN) will be attributed to a value. Several questionnaires offer the options "refused", "legitimate skip", and "don't know". We treated them as missing data. The highest missingness rates for males and females are for the question "Enough money for bills", at 12.4%. We replaced missing values in categorical data with the feature's most frequent value. For numerical data, we used the K-Nearest Neighbors algorithm to impute the data (Cover & Hart, 1967).

The second step involved data processing. We initially applied dummy coding and one-hot encoding to the categorical variables. However, these transformations yielded an AUC score similar to that obtained using the original, unmodified categorical data. We applied Min–Max scaling to normalize the numerical data to the range of 0 to 1, a common approach for non-Gaussian distributions (Han et al., 2012). After imputation, the dataset was split into training (80%) and test (20%) sets for the ML models (excluding the neural network). For the deep learning model (neural network), the dataset was partitioned into training (70%), validation (15%), and test (15%) sets.

More details of the imputation process are found in Supplementary Material, Section 2.1.

[2] https://github.com/rafzgz/ML-add-health-depression

Table 1
Summary of the hyperparameters by model.

Model	Hyperparameters	Notes
Logistic Regression	– $C = 1.0$ (inverse regularization strength) – Solver: 'lbfgs' – $tol = 10^{-4}$ – $max\_iter = 100$	L2 regularization used; solver settings for convergence
Decision Tree	– $min\_samples\_split = 75$ – $min\_samples\_leaf = 150$	Controls tree complexity to reduce overfitting
XGBoost	– max_depth – $learning\_rate$ – $n\_estimators$ – $subsample$ – $colsample\_bytree$	Grid search used; tuned for generalization and overfitting control
SVM	– C (regularization) – γ (for RBF kernel) – Kernel: linear or RBF	Balance between flexibility and generalization
Neural Network	– Hidden layers: 2–6 – Neurons per layer: 32–512 – $learning\_rate$: 10^{-4} to 10^{-2} – Dropout applied	Random search used; final model selected based on validation performance

6.2. Hyperparameters

Hyperparameters are configuration settings that control the learning process of machine learning and deep learning models. Unlike model parameters, which are learned during training, hyperparameters are specified before model fitting and influence key aspects such as model complexity, learning rate, regularization, and architecture. Their selection has a critical impact on model performance and generalizability to unseen data.

We tuned hyperparameters for all five models – logistic regression, decision tree, XGBoost, SVM, and neural network – using grid search or randomized search approaches, depending on the model and computational feasibility. More information about the model's characteristics is provided in the Supplementary material, Sections 2.2 to 2.6. Table 1 has information about the hyperparameter search. We tested every possible combination of the hyperparameters. Final configurations were selected based on performance on a validation set to ensure generalizability. To ensure stability, each configuration was trained 3 times. A detailed explanation of the hyperparameter search is provided in the Supplementary material, Section 2.7.

6.3. Training

All models were trained on the training dataset, with hyperparameters selected via cross-validation or grid search. For the decision tree, XGBoost, and SVMs, we employed 5-fold stratified cross-validation to ensure that each fold preserved the distribution of the outcome classes. This approach helped optimize model performance while mitigating overfitting. Accuracy was monitored during training as a conventional performance indicator, and model comparison.

The neural network was trained using multiple independent runs to ensure stability of results, given the stochastic nature of deep learning. Additionally, during training, the model's performance was evaluated on a held-out validation set after each epoch. This allowed for early stopping and monitoring of generalization performance across epochs, further reducing the risk of overfitting.

6.4. Evaluation and feature importance

The final performance was evaluated on an independent, unseen test set following model training and hyperparameter tuning. This evaluation step is critical for assessing the model's generalization ability to new data and provides an unbiased estimate of predictive performance. We minimized the risk of overfitting by isolating the test set from both training and validation procedures.

We used accuracy as the evaluation metric for the loss function. We applied 1000 bootstrap resamples of the test set for each model to assess the stability and variability of model estimates. This resampling approach enabled us to compute confidence intervals for model accuracy, offering a robust estimate of performance uncertainty. While accuracy provides a general measure of correctness, AUC offers a more informative metric for binary classification by capturing the model's ability to discriminate between classes independently of threshold choice. Therefore, AUC was chosen as the primary performance metric.

To interpret model predictions and assess the relative contribution of input variables, we employed feature importance. In the case of XGBoost, we calculated the feature importance using the 'gain' metric, which refers to the average contribution of each feature to the reduction in the model's objective function, measured across all the trees in the model. This metric highlights the features that most effectively improve the model's predictive accuracy by quantifying their impact on the loss function during training. When interpreting the 'gain' scores, higher values indicate features that play a more significant role in the model's decision-making, providing valuable insights for feature selection and model optimization.

7. Results

7.1. Prediction of depressive symptoms

We evaluated the performance of Logistic Regression, Decision Tree, SVM, XGBoost, and Neural Network models in predicting concurrent adolescent depression symptoms (Wave I → Wave I). We found that XGBoost was the most robust overall performer for both sexes, as shown in Fig. 4. The AUC (top panel) for XGBoost 0.845 (95% CI: 0.825–0.863) was significantly higher than that of the second-best model, SVM (RBF) 0.833 (95% CI: 0.814–0.851), as indicated by a DeLong's test with a z value of 3.735 and a $p < 10^{-3}$. Whereas the Decision Tree had the lowest overall performance, at 0.773 (95% CI: 0.751–0.795). The XGBoost was the most accurate in its predictions (bottom panel), with an accuracy of 0.809 (95% CI: 0.793–0.827). It was followed by SVM (linear kernel) 0.805 (95% CI: 0.788–0.820) and Logistic Regression 0.804 (95% CI: 0.787–0.822). The Decision Tree was the least accurate, with an accuracy of 0.778 (95% CI: 0.759–0.796). Notably, even though the values themselves change, for each model when comparing AUC, the only female model has the best performance, followed by the whole sample, and only males. When comparing XGBoost with the benchmark, Logistic Regression, the improvement was approximately 0.020, with a DeLong's test of ($z = 3.517, p < 10^{-3}$).

Using XGBoost, we show that the top predictors of concurrent adolescent depression mostly reflect how a young person feels about

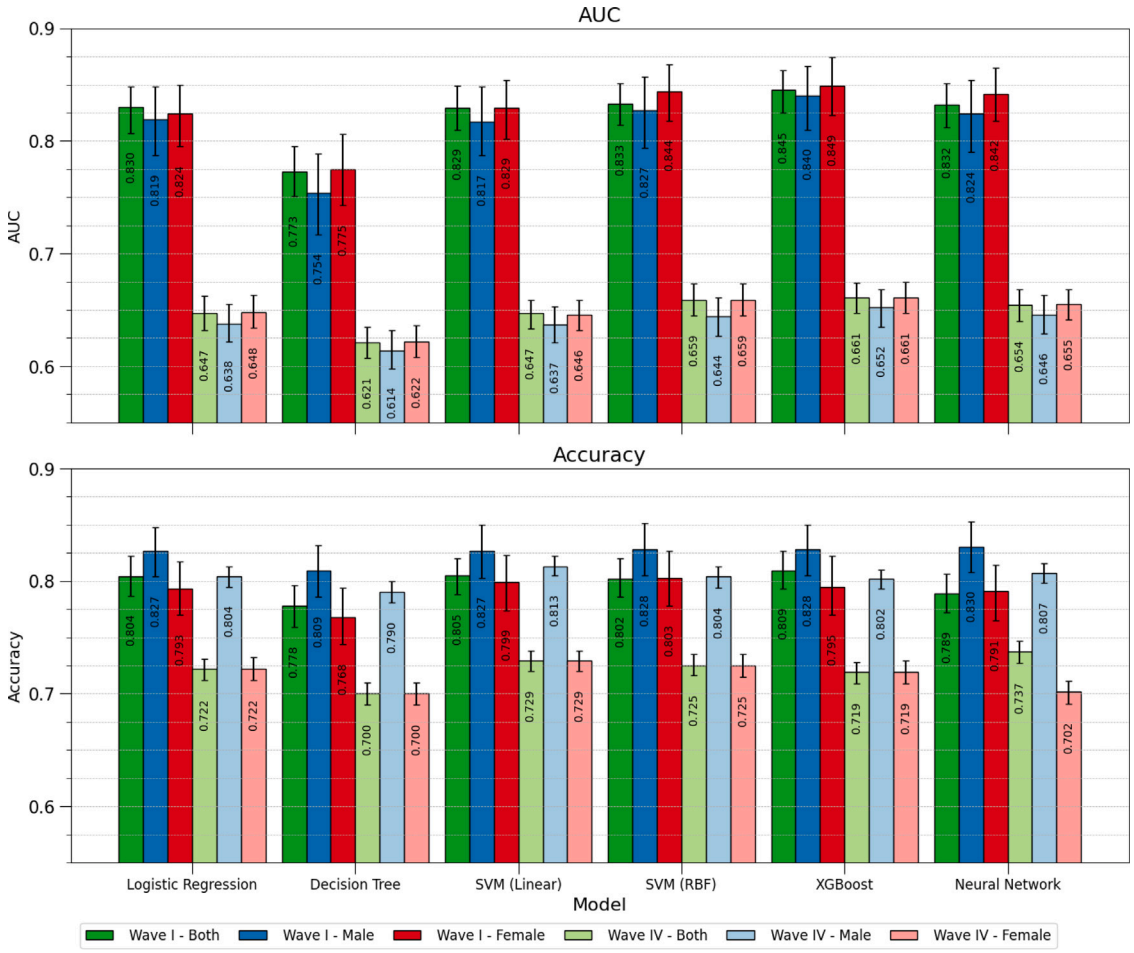


Fig. 4. Comparison of algorithms for predicting depressive symptoms in both the concurrent (Wave I → Wave I) and prospective (Wave I → Wave IV) models, stratified by sex (both sexes, male, and female). The area under the curve (AUC) is presented at the top, while the accuracy metric is shown at the bottom. Error bars were generated using bootstrapping with 1000 iterations, with a 95% confidence interval (CI).

themselves and co-occurring somatic symptoms. Fig. 5 (left) shows the ten most important predictors of depressive symptoms at Wave I, based on their average gain of a feature when it is used in a tree split, in the entire sample. Feeling loved and wanted was the strongest predictor of concurrent adolescent depression (importance score of 85). This was followed by feeling physically weak, liking yourself, and feeling like you are doing everything just right.

The analysis was repeated to predict depression in adulthood, referred to as the prospective modeling (Wave I → Wave IV). Again, XGBoost was the most robust overall performer, as shown in Fig. 4. The AUC was highest for XGBoost 0.661 (95% CI: 0.647–0.674), followed closely by SVM with an RBF kernel 0.659 (95% CI: 0.645–0.673). Logistic Regression also performed competitively, with an AUC of 0.647 (95% CI: 0.632–0.662), while the Decision Tree had the lowest AUC, 0.621 (95% CI: 0.607–0.635). Although the specific values differed, the overall performance trends were consistent across both sexes.

To better understand XGBoost's predictive power, we examined the top-ranked questions based on their average gain importance scores (Fig. 5 - left). Notably, 'Feel love and wanted' was again the most relevant question. Additionally, eight of the ten questions from adolescence also appeared in the depressive symptoms in adulthood.

Additional graphs, including the ROC-AUC curve, confusion matrix, histogram of bootstrap samples, feature importance for other models, and rank concordance between models, are presented in the Supplementary Material, Section 3, Supplementary Figures 2–9.

To prioritize interpretability, we primarily use gain-based feature importance, which provides a straightforward measure of each feature's contribution to model performance. As an additional robustness check, we also computed SHAP (SHapley Additive exPlanations) values for a subset of models (Lundberg & Lee, 2017). SHAP offers a more granular, instance-level decomposition of feature contributions, enabling us to assess both the direction and magnitude of each feature's effect on individual predictions. The detailed SHAP results are available in Supplementary Material, Section 3.1, Supplementary Figures 10 and 11. As expected, strong disagreement with "I feel loved and wanted" and "I do everything just right", and other items in the feeling scale, were associated with an increased predicted risk of depressive symptoms. Similarly, reporting fewer physical health symptoms was associated with a decreased predicted risk of depressive symptoms.

7.2. Predicting clinical diagnoses of depression

We were able to extend the model prediction to accurately classify which participants reported a clinical diagnosis of depression in adolescence (N = 967) and adulthood (N = 1352). XGBoost showed strong predictive performance for both periods, achieving an AUC of 0.788 (95% CI: 0.753–0.818) in adolescence and an AUC of 0.636 (95% CI: 0.602–0.668) in adulthood. We show that despite the temporal gap of over a decade, early-life psychosocial factors retain a meaningful, albeit attenuated, predictive signal for adult depression. Fig. 5 (right) illustrates the most relevant questions for predicting clinical depression, as assessed by self-reports in Wave IV during both adolescence and adulthood. Unlike depressive symptoms, reports of clinical depression showed different predictive factors in adulthood compared to adolescence.

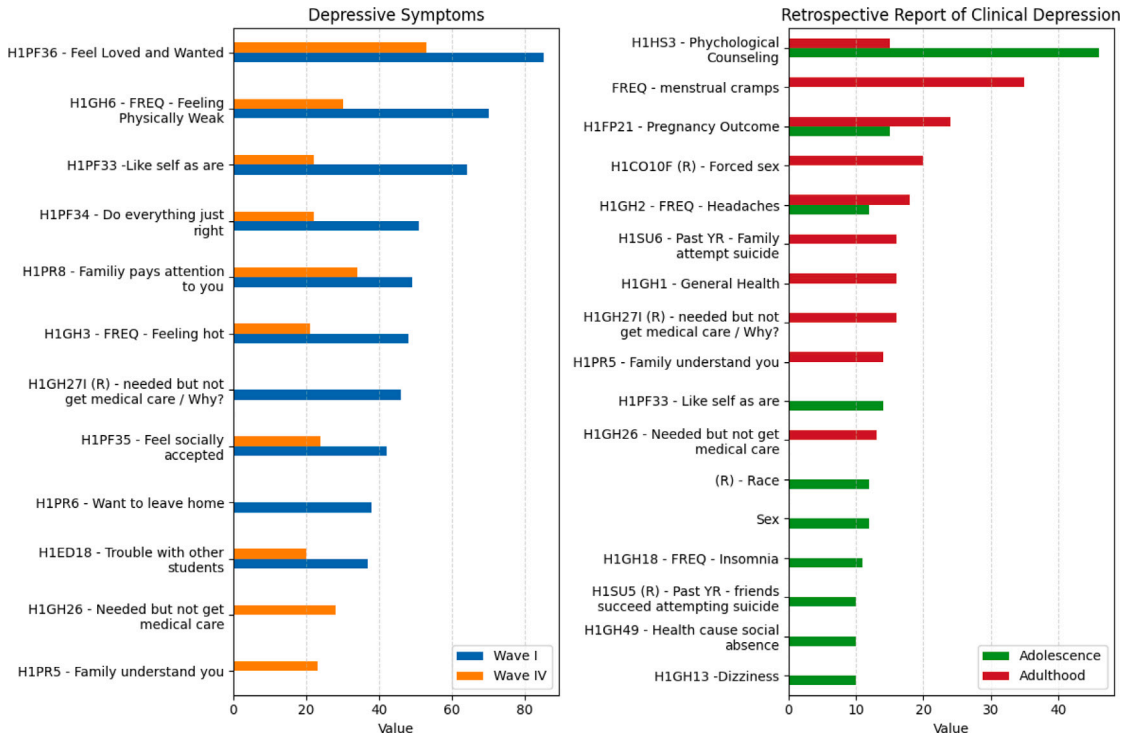


Fig. 5. top 10 feature importance (most important questions) in the XGBoost model: (left) CES-D 19 Wave I and CES-D 10 Wave IV; (right) retrospective report for clinical depression.

7.3. Integrating polygenic scores into depression prediction

We tested three model configurations using XGBoost: (1) environmental factors only, (2) polygenic scores only, and (3) a combined model including both data. Prediction performance was evaluated at two time points: concurrent prediction in adolescence (Wave I → Wave I) and prospective prediction in adulthood (Wave I → Wave IV). We adjusted the environmental input sample by removing entries that did not appear in the other two samples, ensuring a fairer comparison across all inputs. A comparison using the full sample is shown in the Supplementary Material, Section 3.2, Supplementary Figure 12. We found no performance difference in AUC between the environmental-only and combined models, using the whole sample, across Wave I (DeLong's test, $t = 0.47$, $p = 0.6$) and Wave IV (DeLong's test, $t = 0.80$, $p = 0.4$) (Fig. 6). A model based solely on polygenic scores performed only marginally better than chance (AUC ≈ 0.500) at Wave I and even worse than random chance in Wave IV. No meaningful differences were observed between sex-specific models across both Waves.

7.4. Sensitivity analysis

To assess the robustness of the XGBoost model for both sexes, we conducted two additional sensitivity analyses. The results are shown in Table 2. First, we compared the inclusion of all PGS with only the MDD PGS for both Wave I and Wave IV, finding no statistically significant differences. Next, we examined the impact of including or excluding physical and mental health items related to depression. A complete list of the items excluded in each analysis is provided in Supplementary Material, Section 3.3. Excluding physical health items did not significantly affect Wave I, but led to a notable increase in Wave IV. Regarding the exclusion of mental health-related items, removing them significantly reduced predictive power in Wave I, though not as much as anticipated. In Wave IV, exclusion of these items had no impact. Overall, the removal of both physical and mental health items reduced predictive power in both waves. We also conducted a feature importance analysis for models that showed significant differences from the base model, with their corresponding graphs presented in Supplementary Material, Section 3.3, Supplementary Figure 13. Feeling loved and wanted remained the top predictor. The sensitivity analysis feature importance closely mirrored the base model, highlighting the model's consistency and reliability. However, when physical and mental health items were excluded, forced sex emerged as a significant predictor of depressive symptoms.

8. Discussion

This study integrated computational, genetic, and social science perspectives to predict symptom and clinical reports of depression in adolescence and adulthood. We leveraged data from the first Wave of a large nationally representative US longitudinal cohort (Add Health) to compare the performance of several machine learning models, each with increasing levels of complexity: Logistic Regression, Decision Trees, Support Vector Machines, XGBoost, and Neural Networks. XGBoost outperformed all other models across outcomes and subsamples. However, this gain was marginal compared to the benchmark Logistic Regression model (0.02 increase in AUC). We then explored the benefits of the best-fitting model (XGBoost) to determine which factors were most predictive of depressive symptoms. We found that factors that capture feeling loved and accepted, as well as physical stress indicators, were the strongest predictors of depressive symptoms in both adolescence and adulthood. Different factors emerged as the most influential for reporting a clinical diagnosis, with stressful or traumatic life events being the most predictive. However, this relationship differed between adolescents and adults. In adulthood, female-specific factors and trauma, such as forced sex and family member attempted suicide, were most predictive, whereas in adolescence, sex, race, and trauma were most predictive. In both cases, psychological counseling was the most predictive. As such life-changing and traumatic events often precede the onset of depression, these findings highlight the importance of regular and ongoing mental health assessments, facilitated by universal access to affordable, high-quality care across the life course. Furthermore, including genetic risk indexed by multiple

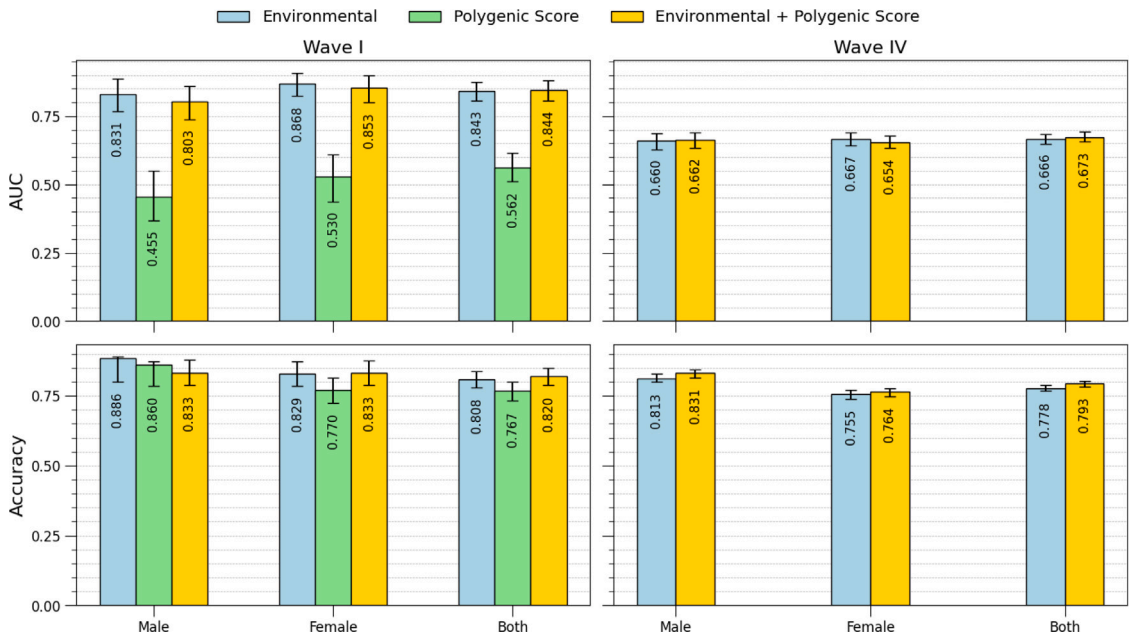


Fig. 6. The AUC for the 18 different input configurations. The graph does not display the polygenic score for Wave IV because the results are worse than those expected by chance. The error bars were created using Bootstrapping for $N = 1000$, with a 95% confidence interval (CI). The environmental data is reduced to match the sample size of the polygenic score, and the environmental + polygenic score used $N_{PGS} = 5731$.

Table 2
Sensitivity analysis.

Wave	Sensitivity	AUC	Base	AUC	deLong Test
I	MDD PGS	0.835 (0.799–0.869)	All PGS	0.844 (0.811–0.881)	$z = 1.62, p = 0.1$
IV	MDD PGS	0.667 (0.649–0.684)	All PGS	0.673 (0.634–0.689)	$z = 0.40, p = 0.7$
I	Excluded Physical	0.844 (0.826–0.863)	All Envi	0.845 (0.825–0.863)	$z = 0.81, p = 0.4$
IV	Excluded Physical	0.666 (0.655–0.676)	All Envi	0.661 (0.647–0.674)	$z = 2.58, p < 0.05$
I	Excluded Mental	0.838 (0.819–0.857)	All Envi	0.845 (0.825–0.863)	$z = 3.20, p < 0.05$
IV	Excluded Mental	0.666 (0.656–0.677)	All Envi	0.661 (0.647–0.674)	$z = -0.35, p = 0.7$
I	Excluded M. and P.	0.828 (0.808–0.847)	All Envi	0.845 (0.825–0.863)	$z = 3.39, p < 0.05$
IV	Excluded M. and P.	0.656 (0.645–0.666)	All Envi	0.661 (0.647–0.674)	$z = 4.19, p < 0.05$

PGS did not significantly improve model fit for depressive symptoms both concurrently and longitudinally. This suggests that while genetic factors play a role in depression, they do not add substantial predictive value beyond the environmental and psychological factors captured in our models. The top predictors remained consistent across sensitivity analyses, whereby we removed sets of physical and mental health-related items, with 'feeling loved and wanted' continuing to be the top predictor. However, 'forced sex' then emerged as a significant predictor in the depressive symptoms model in both adolescence and adulthood. These results demonstrate the stability of our model and the importance of psychosocial factors in predicting depressive symptoms, while also highlighting the nuanced role of certain life experiences in shaping these outcomes.

To benchmark the value added of our study, we compared the AUC values obtained in our study with those from a prior research that utilized Add Health data to predict short-term adolescent depression. This earlier study employed boosted classification and regression techniques to forecast depression one year later using a 20-item index, achieving an AUC of approximately 0.80 (Voorhees et al., 2008). We also compared our results with those of (Lu et al., 2023), which reported an accuracy of about 0.68 for predicting major depression, compared to 0.66 with environmental data alone and 0.63 with PGS alone. In comparison, our models demonstrated higher predictive accuracy, suggesting that incorporating a broader set of predictors and leveraging more complex machine learning methods can enhance the identification of adolescents at risk of depression. Moreover, we observed statistically significant, though modest, gains over the benchmark logistic regression commonly used in the social sciences, consistent with improvements reported in other studies. Some reports suggest that XGBoost is suitable for large cohort studies on mental health (Liu et al., 2023; Sharma & Verbeke, 2020; Smith et al., 2025). However, others have demonstrated that ML approaches, such as SVM, do not provide a substantial benefit in model fit over traditional regression approaches when applied to predict schizophrenia (Bracher-Smith et al., 2022). Due to the small sample sizes available in longitudinal, epidemiological cohorts like Add Health, our findings caution against universally adopting ML methods in these datasets just yet, since logistic regression currently performs about as well as our ML methods. However, one potential benefit of ML methods is the ability to obtain importance scores, which can provide insights into predictor variables. With more sophisticated methods or larger sample sizes in the future, the utility of these models, including their interpretability, will likely significantly improve.

We demonstrate that using key indicators of depression risk, logistic regression is nearly as effective as more complex ML models. However, to achieve optimal predictive performance, more sophisticated ML approaches are necessary. This highlights both the applicability and feasibility of complex ML models for capturing multifactorial patterns of depression risk in large cohort datasets. This is supported by recent advances in ML and artificial intelligence research, which are reshaping the landscape of psychiatric research and clinical practice (Chen et al., 2022). The individual-level prediction demonstrated in this study enables a more individualized and data-driven approach to diagnosis. A systematic review (Meehan et al., 2022) found that prediction models are often biased, demonstrate overfitting, and have limited generalizability. Here, we aimed to reduce bias by employing

multiple machine learning methods and various model configurations. To mitigate overfitting, we utilized cross-validation and regularization techniques. Finally, we used a nationally representative cohort to maximize the generalizability to the US population. Overall, applying machine learning models to large cohorts is necessary to achieve an optimal prediction of self-reported depressive symptoms.

The most influential factors in this study for predicting self-report depressive symptoms in both adolescence and adulthood were self-perception and general health. These findings align with previous research, as shown in studies on stigma's impact on health (Hatzenbuehler et al., 2013), cognitive therapy's effectiveness in treating mood disorders (Beck, 2005), and psychological interventions for depression prevention in youth (Hetrick et al., 2016). Studies show that a young person's self-perception is a key contributor to well-being during adolescence and is associated with improved long-term health outcomes (Hoyt et al., 2012). Somatic symptoms during adolescence can independently predict severe mental illness in adulthood, even in the absence of diagnosed anxiety or depression (Bohman et al., 2018). Extensive research has demonstrated a strong link between physical and mental health, with conditions such as poor general health often co-occurring with depressive symptoms (Fiorillo et al., 2023). Childhood adversity is another potent risk factor for adult depression, with emotional abuse, neglect, and other traumas increasing vulnerability through cumulative exposure to stressors over the life course (Korkeila et al., 2010). Such adversities not only elevate the risk of later depression but also increase the likelihood of experiencing additional adverse events in adulthood, increasing mental health challenges. These findings suggest that interventions addressing both self-perception and general health from an early age could play a critical role in preventing depressive symptoms across the lifespan, particularly in the face of childhood adversity and physical health challenges.

Traumatic and life-altering events were most predictive for self-reports of a depression diagnosis from a clinician, which differed from our findings based on depressive symptom counts from the CES-D. This supports the notion that early trauma can contribute to the development of depression in adulthood (Mandelli et al., 2015; Negele et al., 2015). The distinction between self-reports of a depression diagnosis and depressive symptoms in our findings likely stems from differences in both severity and the diagnostic criteria used to define each condition. Clinical depression, typically diagnosed through standardized diagnostic criteria, involves the same symptoms as self-reported depressive symptoms, but with the additional requirement of symptom persistence (lasting at least two weeks) and a substantial impact on daily functioning. These symptoms are often more severe, with individuals generally seeking medical attention only when the depression significantly disrupts their ability to function. Clinical depression is also frequently linked to significant life events, such as trauma or other life-altering experiences, which can profoundly affect an individual's psychological functioning and contribute to the onset and exacerbation of depressive episodes (Mandelli et al., 2015). In contrast, self-reported depressive symptoms tend to capture a broader spectrum of emotional distress, which changes day to day and may not meet the clinical threshold for a formal diagnosis. As such, while trauma and life-altering events are key factors in clinical depression, they may not be as relevant in predicting self-reported depressive symptoms, where self-perception often emerges as a more potent predictor.

Many of the top predictors in the model, such as "feeling loved and wanted" and "like self as are", reflect aspects of emotional and social well-being that are closely intertwined with depression (Cohen & Wills, 1985; Orth et al., 2008). These factors are known to be highly correlated with depressive symptoms and could function as early indicators of mental health distress. Many studies have shown the association between depression and constructs like self-esteem, social connectedness, and perceived support from family and peers (Keane & Loades, 2017). By incorporating such predictors, we aimed to cast a wide net and capture the multifaceted nature of adolescent well-being, recognizing that the onset of depression may not always present with clearly defined clinical symptoms early on. Furthermore, factors closely related to depression may serve as modifiable targets for early interventions, offering a more nuanced approach to identifying adolescents at risk as they provide insights into the context of depressive symptomatology, which may not be fully captured by clinical indicators alone.

The addition of PGS did not substantially improve the prediction of depressive symptoms. During adolescence, the prediction gain was not significant, whereas in adulthood, the increase was marginal. This was even the case when only including the polygenic scores for major depressive disorder. We emphasize that genetic influence has a modest effect in comparison with the environment. These gains are significant but may not be particularly useful in identifying who is most at risk. Similar findings were also reported from a review of longitudinal risk factors for depression, highlighting the role of social support, physical health comorbidities, and community engagement (de Sousa et al., 2025). This limitation is not exclusive to depression, but has also been reported to have limited predictive value for well-being outcomes, as it did not improve the prediction model beyond environmental and psychosocial factors (Pelt et al., 2024). The specific exposome, which encompasses factors such as personality, optimism, and social support, was by far the most predictive of well-being.

This work is subject to several limitations. First, although XGBoost achieves strong performance, it is less interpretable than traditional statistical models, which complicates insight into causal mechanisms. Additionally, the use of XGBoost's gain-based feature importance may be biased by correlated predictors, potentially distorting the actual contributions of individual variables. Second, the temporal distance between predictors and long-term outcomes introduces substantial noise due to unmeasured life events, shifting environments, and personal experiences. Third, all diagnoses were self-reported and may be influenced by differential access to healthcare, stigma, or recall bias. Fourth, the absence of improvement from polygenic scores may reflect current limitations in their construction, transferability, or predictive relevance across diverse populations, as well as the sample size and power when working with PGS. It is worth noting that we were only using European-like ancestries when working with genetic data, which limits the generalizability of our findings to other ancestries. Finally, while machine learning models have shown promise in predicting clinical outcomes, their application in routine clinical practice is premature.

This study demonstrates the potential of machine learning models to predict depression using comprehensive datasets, such as Add Health, which encompass both environmental and genetic factors. However, translating these findings into clinical practice requires careful consideration, as depression in real-world settings is primarily diagnosed through symptom assessments, medical records, and self-reports (e.g., PHQ-9 or CES-D). As highlighted by Meehan and Danese (2021), challenges such as generalizability, interpretability, and clinical utility must be addressed before machine learning models can be reliably implemented in healthcare settings. While our models perform well with rich, multidimensional data, they must be adapted to the more limited, symptom-based data typically available in clinical settings. To bridge this gap, future research could focus on developing a model that enhances the feasibility of applying this methodology in clinical settings. A limitation of the current approach is the requirement for new patients to complete a full set of 72 or 75 survey questions (for males and females), which may not be practical in clinical environments due to time constraints and patient burden. To address this, one possible solution would be to reduce the number of questions by applying a "relevance threshold" to select for those with higher feature importance, while removing those deemed to have low relevance. Additionally, PGS could be excluded, as they are not fit for clinical settings and did not substantially improve model performance. By streamlining the

questionnaire through this targeted selection process, the model could become more feasible for clinical implementation.

Another potential avenue for extending future research could explore the application of these models to larger cohorts, such as the UK Biobank. This approach would allow for non-linear interactions between a greater number of variables, potentially yielding more accurate predictions and highlighting a bigger gap between Logistic Regression and more complex models. Although biological data did not enhance prediction in this study, combining psychosocial data with larger datasets may reveal more complex relationships and improve the understanding of depression's etiology. Additionally, incorporating time-to-event or age-of-onset analyses could help capture temporal dynamics and distinguish short- from long-term risk trajectories of depression.

For social scientists working with smaller datasets, traditional methods like Logistic Regression remain valuable for capturing key predictors without the complexity of machine learning models. Moving forward, advancing life-course models of mental health risk will depend on leveraging larger and more diverse datasets to clarify how psychosocial and environmental factors influence depression over time, with the aim of creating practical tools for early detection and prevention.

In conclusion, our findings underscore the predictive strength of psychosocial data collected in adolescence for identifying individuals at risk of depression, particularly in the short term. However, predicting long-term depression in adulthood remains more challenging. For depressive symptoms, the prediction for adolescents and adults aligns with the most important factors, such as self-perception and general health; this is not the case for clinical depression, where trauma and life-altering events play a larger role. Since our top findings align with existing literature on the predictors of depression, it is also possible that lower-ranked predictors may show meaningful correlations with depression diagnoses and warrant further investigation. Furthermore, the lack of added value from polygenic scores in this study suggests that, for early screening efforts, prioritizing contextual and environmental data may be more effective than genetic factors.

CRediT authorship contribution statement

Rafael Geurgas: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Saul J. Newman:** Writing – review & editing. **Evelina T. Akimova:** Writing – review & editing, Writing – original draft, Formal analysis. **Katherine N. Thompson:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Robbee Wedow:** Writing – review & editing, Project administration, Formal analysis.

Ethical statement

Use of the Add Health data at Purdue University was approved by the Purdue Institutional Review Board (IRB) under application IRB-2023-1819.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Robbee Wedow is a research fellow at AnalytiXIN, which is a consortium of health-data organizations, industry partners, and university partners in Indiana, primarily funded through the Lilly Endowment, IU Health, and Eli Lilly and Company.

Acknowledgments

This research uses data from Add Health, funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. The Add Health phenotypic data was applied for and approved under restricted-use application #27111802. The Add Health genetic data was applied for and approved under application #38868 using the National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI)'s database of Genotypes and Phenotypes (dbGaP). Saul Newman and this project were supported by UKRI project FINDME (EP/Y023080/1). Katherine Thompson and Robbee Wedow equally contributed for project supervision.

Appendix A. Supplementary material

Supplementary methods and data related to this article can be found online at https://doi.org/10.1016/j.ssmph.2025.101886.

Data availability

The authors do not have permission to share data.

References

Ahmed, R., Bathina, K., Fazel, M., & Butte, A. (2019). Artificial intelligence techniques for detection of depression: A review. *Current Psychiatry Reports*, *21*(10), 108.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders, fifth edition (DSM-5)*. Arlington, VA: American Psychiatric Publishing.

Appleby, L., Hunt, I. M., & Kapur, N. (2017). New policy and evidence on suicide prevention. *The Lancet Psychiatry*, *4*(9), 658–660.

Backes, E. P., & Bonnie, R. J. (2019). The promise of adolescence: Realizing opportunity for all youth.

Bader, M., Abdelwanis, M., Maalouf, M., & Jelinek, H. F. (2024). Detecting depression severity using weighted random forest and oxidative stress biomarkers. *Scientific Reports*, *14*(1), 16328.

Beck, A. T. (2005). The current state of cognitive therapy: a 40-year retrospective. *Archives of General Psychiatry*, *62*(9), 953–959.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561–571.

Boateng, E. Y., Abaye, D. A., et al. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, *7*(04), 190.

Bohman, H., Låftman, S. B., Cleland, N., Lundberg, M., Päären, A., & Jonsson, U. (2018). Somatic symptoms in adolescence as a predictor of severe mental illness in adulthood: a long-term community-based follow-up study. *Child and Adolescent Psychiatry and Mental Health*, *12*, 1–12.

Bracher-Smith, M., Rees, E., Menzies, G., Walters, J. T., O'Donovan, M. C., Owen, M. J., Kirov, G., & Escott-Price, V. (2022). Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. *Schizophrenia Research*, *246*, 156–164.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Brown, G. W., & Harris, T. (2012). *Social origins of depression: A study of psychiatric disorder in women*. Routledge.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). *Statistics versus machine learning*: *Vol. 15*, Nat. Meth..

Chaudhury, M. D., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. Microsoft Research White Paper.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., & et al. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250.

Chen, Z. S., Galatzer-Levy, I. R., Bigio, B., Nasca, C., Zhang, Y., et al. (2022). Modern views of machine learning for precision psychiatry. *Patterns*, *3*(11).

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chung, W.-Y., Chen, Y.-H., & Lin, Y.-L. (2020). Mental health prediction using machine learning: Taxonomy, applications, and challenges. *Computer Methods and Programs in Biomedicine*, *196*, Article 105608.

Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, *98*(2), 310.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Development and natural history of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*(8), 837–844.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *13*(1), 21–27.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *20*(2), 215–232.

de Mello, M. F., de Jesus Mari, J., Bacaltchuk, J., Verdeli, H., & Neugebauer, R. (2005). A systematic review of research findings on the efficacy of interpersonal therapy for depressive disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *255*(2), 75–82.

de Sousa, R. D., Zagalo, D. M., Costa, T., de Almeida, J. M. C., Canhão, H., & Rodrigues, A. (2025). Exploring depression in adults over a decade: a review of longitudinal studies. *BMC Psychiatry*, *25*(1), 378.

Del Pozo-Banos, M., John, A., Petkov, N., et al. (2021). Predicting and diagnosing depression using electronic health records: A systematic review. *Journal of Affective Disorders*, *273*, 391–400.

Dennis, A. C., Will, S., Harris, K. M., & Hummer, R. A. (2022). Depressive symptoms in the national longitudinal study of adolescent to adult health (add health). *Carolina Digital Repository*.

Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. *Science*, *196*(4286), 129–136.

Ezzy, D. (1993). Unemployment and mental health: A critical review. *Social Science and Medicine*, *37*(1), 41–52. http://dx.doi.org/10.1016/0277-9536(93)90316-V.

Feldman, G. (2007). Cognitive and behavioral therapies for depression: overview, new directions, and practical recommendations for dissemination. *Psychiatric Clinics of North America*, *30*(1), 39–50.

Fiorillo, A., de Girolamo, G., Simunovic, I. F., Gureje, O., Isaac, M., Lloyd, C., Mari, J., Patel, V., Reif, A., Starostina, E., et al. (2023). The relationship between physical and mental health: an update from the WPA working group on managing comorbidity of mental and physical health. *World Psychiatry*, *22*(1), 169.

Frasquilho, D., Matos, M. G., Salonna, F., Guerreiro, D., Storti, C. C., Gaspar, T., & Caldas-de Almeida, J. M. (2015). Mental health outcomes in times of economic recession: a systematic literature review. *BMC Public Health*, *16*(1), 115.

Frick, A., Howner, K., Fischer, H., Kristiansson, M., & Furmark, T. (2014). Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A multimodal machine learning study. *Brain and Behavior*, *4*(3), 337–350.

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 1–11.

Fried, E. I., & Nesse, R. M. (2017). The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197.

Fryers, T., Melzer, D., & Jenkins, R. (2003). Social inequalities and the common mental disorders. *Social Psychiatry and Psychiatric Epidemiology*, *38*(5), 229–237. http://dx.doi.org/10.1007/s00127-003-0627-2.

Gao, S., Calhoun, V. D., & Sui, J. (2020). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *26*(7), 584–598.

Gitterman, A. (1991). *Handbook of social work practice with vulnerable populations*. Columbia University Press.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Waltham: Morgan Kaufmann Publishers.

Harris, K. M., Halpern, C. T., Whitsel, E. A., Hussey, J. M., Killeya-Jones, L. A., Tabor, J., & Dean, S. C. (2019). Cohort profile: The national longitudinal study of adolescent to adult health (add health). *International Journal of Epidemiology*, *48*(5), 1415–1415k.

Hatzenbuehler, M. L., Phelan, J. C., & Link, B. G. (2013). Stigma as a fundamental cause of population health inequalities. *American Journal of Public Health*, *103*(5), 813–821.

Hetrick, S. E., Cox, G. R., Witt, K. G., Bir, J. J., & Merry, S. N. (2016). Cognitive behavioural therapy (CBT), third-wave CBT and interpersonal therapy (IPT) based interventions for preventing depression in children and adolescents. *Cochrane Database of Systematic Reviews*, (8).

Hirschfeld, R. M. (2012). The epidemiology of depression and the evolution of treatment. *The Primary Care Companion for CNS Disorders*, *14*(Suppl. 1: Editor Choice), 26328.

Hirschfeld, G., Mehler, D. M. A., & Hagemann, D. (2022). Identifying depression through machine learning analysis of omics data: Scoping review. *JMIR Mental Health*, *9*(6), Article e35615.

Horwitz, A. V. (2007). Transforming normality into pathology: the DSM and the outcomes of stressful social arrangements. *Journal of Health and Social Behavior*, *48*(3), 211–222.

Hoyt, L. T., Chase-Lansdale, P. L., McDade, T. W., & Adam, E. K. (2012). Positive youth, healthy adults: does positive well-being in adolescence predict better perceived health and fewer risky health behaviors in young adulthood? *Journal of Adolescent Health*, *50*(1), 66–73.

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455–2465.

Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare*: *Vol. 11*, (p. 285). MDPI.

Jacobson, N. S., & Addis, M. E. (1993). Research on couples and couple therapy: What do we know? Where are we going? *Journal of Consulting and Clinical Psychology*, *61*(1), 85.

Jahoda, M. (1988). Economic recession and mental health: Some conceptual issues. *Journal of Social Issues*, *44*(4), 13–23.

Jansson, M., Gatz, M., Berg, S., Johansson, B., Malmberg, B., McClearn, G. E., Schalling, M., & Pedersen, N. L. (2004). Gender differences in heritability of depressive symptoms in the elderly. *Psychological Medicine*, *34*, 471–479. http://dx.doi.org/10.1017/S0033291703001375.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Keane, L., & Loades, M. (2017). Low self-esteem and internalizing disorders in young people–a systematic review. *Child and Adolescent Mental Health*, *22*(1), 4–15.

Kessing, L. V., Ziersen, S. C., Caspi, A., Moffitt, T. E., & Andersen, P. K. (2023). Lifetime incidence of treated mental health disorders and psychotropic drug prescriptions and associated socioeconomic functioning. *JAMA Psychiatry*, *80*(10), 1000–1008.

Kessler, R. C., Berglund, P., Demler, O., & et al. (2003). The epidemiology of major depressive disorder: results from the national comorbidity survey replication (NCS-R). *JAMA*, *289*(23), 3095–3105.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 593–602.

Kessler, R. C., & Essex, M. (1982). Marital status and depression: The importance of coping resources. *Social Forces*, *61*(2), 484–507.

Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539–548. http://dx.doi.org/10.1037/0022-006X.73.3.539.

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., et al. (2024). The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58–90.

Korkeila, J., Vahtera, J., Nabi, H., Kivimäki, M., Korkeila, K., Sumanen, M., Koskenvuo, K., & Koskenvuo, M. (2010). Childhood adversities, adulthood life events and depression. *Journal of Affective Disorders*, *127*(1–3), 130–138.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, J. (2011). Pathways from education to depression. *Journal of Cross-Cultural Gerontology*, *26*(2), 121–135.

Liu, D., Chen, Z., Marrero, W. J., Jacobson, N. C., & Thesen, T. (2023). Explainable machine learning-based prediction of depression severity in medical students.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, *380*(9859), 2095–2128. http://dx.doi.org/10.1016/S0140-6736(12)61728-0.

Lu, T., Silveira, P. P., & Greenwood, C. M. (2023). Development of risk prediction models for depression combining genetic and early life risk factors. *Frontiers in Neuroscience*, *17*, Article 1143496.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mabry, P. L., Olster, D. H., Morgan, G. D., & Abrams, D. B. (2008). Interdisciplinarity and systems science to improve population health: a view from the NIH office of behavioral and social sciences research. *American Journal of Preventive Medicine*, *35*(2), S211–S224.

Mandelli, L., Petrelli, C., & Serretti, A. (2015). The role of specific early trauma in adult depression: A meta-analysis of published literature. Childhood trauma and adult depression. *European Psychiatry*, *30*(6), 665–680.

Matheson, F. I., Moineddin, R., Dunn, J. R., Creatore, M. I., Gozdyra, P., & Glazier, R. H. (2006). Urban neighborhoods, chronic stress, gender and depression. *Social Science & Medicine*, *63*(10), 2604–2616.

McGue, M., & Christensen, K. (2003). The heritability of depression symptoms in elderly Danish twins: Ocassion-specific versus general effects. *Behavior Genetics*, *33*(2), 83–93. http://dx.doi.org/10.1023/A:1022545600034.

McPherson, S., & Armstrong, D. (2006). Social determinants of diagnostic labels in depression. *Social Science & Medicine*, *62*(1), 50–58.
Meehan, A. J., & Danese, A. (2021). Progress and barriers to the implementation of prediction modelling in child and adolescent mental health—A commentary on senior et al. *JCPP Advances*, *1*(4), Article e12052.
Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*(6), 2700–2708.
Mills, K. L., Lalonde, F., Clasen, L. S., Giedd, J. N., & Blakemore, S.-J. (2014). Developmental changes in the structure of the social brain in late childhood and adolescence. *Social Cognitive and Affective Neuroscience*, *9*(1), 123–131.
Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, *13*, 23–47.
Monroe, S. M., & Simons, A. D. (2005). Understanding the course and etiology of depression: Stressful life events and vulnerabilities to depression. *Annual Review of Psychology*, *56*, 17–37.
Musliner, K. L., Seifuddin, F., Judy, J. A., Pirooznia, M., Goes, F. S., & Zandi, P. P. (2015). Polygenic risk, stressful life events and depressive symptoms in older adults: a polygenic score analysis. *Psychological Medicine*, *45*(8), 1709–1720.
Nathan, P. E. (2007). Efficacy, effectiveness, and the clinical utility of psychotherapy research. *The Art and Science of Psychotherapy*, 69–83.
Negele, A., Kaufhold, J., Kallenbach, L., & Leuzinger-Bohleber, M. (2015). Childhood trauma and its relation to chronic depression in adulthood. *Depression Research and Treatment*, *2015*(1), Article 650804.
Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low self-esteem prospectively predicts depression in adolescence and young adulthood. *Journal of Personality and Social Psychology*, *95*(3), 695.
Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: a global public-health challenge. *The Lancet*, *369*(9569), 1302–1313.
Pearlin, L. I., & Johnson, J. S. (1977). Marital status, life-strains and depression. *American Sociological Review*, 704–715.
Pelt, D. H., Habets, P. C., Vinkers, C. H., Ligthart, L., van Beijsterveldt, C. E., Pool, R., & Bartels, M. (2024). Building machine learning prediction models for well-being using predictors from the exposome and genome in a population cohort. *Nature Mental Health*, *2*(10), 1217–1230.
Peterson, C. (2009). Psychological approaches to mental illness. In T. L. Scheid, & T. N. Brown (Eds.), *A handbook for the study of mental health: social contexts, theories, and systems* (2nd ed.). (pp. 89–105). Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511984945.008.
Plomin, R., & Von Stumm, S. (2022). Polygenic scores: prediction versus explanation. *Molecular Psychiatry*, *27*(1), 49–52.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.
Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.
Reading, R., & Reynolds, S. (2001). Debt, social disadvantage and maternal depression. *Social Science & Medicine*, *53*(4), 441–453.
Remes, O., Mendes, J. F., & Templeton, P. (2021). Biological, psychological, and social determinants of depression: a review of recent literature. *Brain Sciences*, *11*(12), 1633.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).
Richter, T., Fishbain, B., Richter-Levin, G., & Okon-Singer, H. (2021). Machine learning-based behavioral diagnostic tools for depression: advances, challenges, and future directions. *Journal of Personalized Medicine*, *11*(10), 957.
Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., et al. (2019). Integrating the hierarchical taxonomy of psychopathology (HiTOP) into clinical practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.
Sau, A., & Bhakta, I. (2019). Machine learning algorithms for predicting depression and anxiety: A systematic review. *Current Psychiatry Reports*, *21*(7), 44.
Schwartz, S., & Corcoran, C. (2009). Biological theories of psychiatric disorders: A sociological approach. In T. L. Scheid, & T. N. Brown (Eds.), *A handbook for the study of mental health: social contexts, theories, and systems* (2nd ed.). (pp. 64–88). Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511984945.007.
Sharma, A., & Verbeke, W. J. (2020). Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers dutch dataset (n=11,081). *Frontiers in Big Data*, *3*, Article 523466.
Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448.
Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist*, *65*(2), 98.
Slavich, G. M., & Irwin, M. R. (2014). From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. *Psychological Bulletin*, *140*(3), 774.
Smith, A., Miller, J. J., Anthony, D. C., & Radford-Smith, D. E. (2025). Machine learning-based prediction of anxiety disorders using blood metabolite and social trait data from the UK biobank. *Brain, Behavior, & Immunity-Health*, Article 101010.
Solmi, M., Cortese, S., Vita, G., De Prisco, M., Radua, J., Dragioti, E., Köhler-Forsberg, O., Madsen, N. M., Rohde, C., Eudave, L., et al. (2023). An umbrella review of candidate predictors of response, remission, recovery, and relapse across mental disorders. *Molecular Psychiatry*, *28*(9), 3671–3687.
Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Informatics Research*, *16*(4), 253–259.
Stetler, C., & Miller, G. E. (2011). Depression and hypothalamic-pituitary-adrenal activation: a quantitative summary of four decades of research. *Biopsychosocial Science and Medicine*, *73*(2), 114–126.
Sullivan, P., Neale, M., & Kendler, K. (2000). Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry*, *157*(10), 1552–1562. http://dx.doi.org/10.1176/appi.ajp.157.10.1552.
Thompson, K. N., Couvy-Duchesne, B., Lee, S. H., Geurgas, R., Jeong, Y., Arirangan, S., & Tropf, F. C. (2024). Illuminating the complex interplay of risk factors for depression within a large-scale US longitudinal cohort. Retrieved from https://osf.io/preprints/psyarxiv/wf52a_v1.
Unnikrishnan, P., Kumar, D. K., Poosapadi Arjunan, S., Kumar, H., Mitchell, P., & Kawasaki, R. (2016). Development of health parameter model for risk prediction of CVD using SVM. *Computational and Mathematical Methods in Medicine*, *2016*(1), Article 3016245.
Van Voorhees, B. W., Paunesku, D., Gollan, J., Kuwabara, S., Reinecke, M., & Basu, A. (2008). Predicting future risk of depressive episode in adolescents: the chicago adolescent depression risk assessment (CADRA). *The Annals of Family Medicine*, *6*(6), 503–511.
Verhagen, M. D. (2023). Incorporating machine learning into sociological model-building. *Sociological Methodology*, Article 00811750231217734.
Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., et al. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries: a systematic analysis for the global burden of disease study 2013. *The Lancet*, *386*(9995), 743–800. http://dx.doi.org/10.1016/S0140-6736(15)60692-4.
World Health Organization (2023). Depression. https://www.who.int/news-room/fact-sheets/detail/depression.
Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*(5), 668–681. http://dx.doi.org/10.1038/s41588-018-0090-3.
Xin, Y., & Ren, X. (2022). Predicting depression among rural and urban disabled elderly in China using a random forest classifier. *BMC Psychiatry*, *22*(1), 118.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.