



A Multimodal Ensemble Learning Framework for Early Detection of Adolescent Depression Using Multi-Source Real-World Data

Ramesh K. Bhukya^{1,*}, Nirbhay Singh² and Aniket Tiwari³

¹Department of Electronics and Communication Engineering, Indian Institute of Information Technology Allahabad, Prayagraj 211015, India

²Department of Electronics and Communication Engineering, Guru Ghasidas Vishwavidyalaya, Bilaspur 495009, India

³Department of Computer Science and Engineering, United College of Engineering and Research, Prayagraj 211010, India

Abstract

Depression and academic stress undermine students' attendance, performance, and retention, yet educational institutions lack reliable early-alert tools. Because about three-quarters of lifetime psychiatric conditions emerge before age 25, timely automated screening is critical for early intervention. This paper proposes a multi-source ensemble learning framework for distinguishing *depression* (DP) from *non-depression* (ND) using 1,547 harmonized samples drawn from six real-world sources: occupational surveys, student academic stress records, Twitter sentiment data, Reddit depression posts, fine-grained emotion corpora, and behavioural monitoring signals. Academic stress and behavioural records serve as the core educational sources, while the remaining sources broaden linguistic and affective coverage. Six *machine learning* (ML) classifiers were benchmarked

under 10-fold cross-validation: *decision tree* (DT), *support vector machine* (SVM), *random forest* (RF), *adaptive boosting* (AdaBoost), *extreme gradient boosting* (XGBoost), and *light gradient boosting machine* (LightGBM). RF performed best, achieving 97.996% accuracy, the highest sensitivity (96.741%), only 6 false positives, and 25 missed DP cases. LightGBM recorded the highest specificity (99.615%), making it suitable for precision-oriented screening, whereas the boosting ensembles performed comparably to the linear SVM and all ensembles clearly outperformed the DT baseline. These findings support ensemble learning, particularly bagging, for heterogeneous, multi-source student mental health data and suggest practical value for counsellors and student support services in identifying at-risk students early.

Keywords: educational data mining, student depression, academic stress, student well-being, learning analytics, early warning systems.

1 Introduction

Psychological well-being challenges and emotional distress are among the major causes of disability and



Submitted: 27 August 2026

Revised: 21 September 2026

Accepted: 28 September 2026

Published: 01 October 2026

Vol. 2, No. 2, 2026.

10.62762/TEDM.2026.244050

*Corresponding author:

✉ Ramesh K. Bhukya

rkbhukya@iiita.ac.in

Citation

Bhukya, R. B., Singh, N., & Tiwari, A. (2026). A Multimodal Ensemble Learning Framework for Early Detection of Adolescent Depression Using Multi-Source Real-World Data. *ICCK Transactions on Educational Data Mining*, 2(2), 63–84.

© 2026 ICCK (Institute of Central Computation and Knowledge)

disease burden globally, with onset disproportionately occurring during adolescence and early adulthood. Large epidemiological studies indicate that approximately three-quarters of all lifetime psychological conditions emerge before the age of 25. This makes youth a critical period for development, prevention, early-stage identification, and intervention [1, 2]. Conditions that develop during this timeframe are strongly linked to subsequent academic difficulties, occupational instability, and higher risks of severe psychological crises that often extend long into adulthood. Before 2010, youth mental health care was marked by a diagnosis-based clinical paradigm that was adult-centric. The diagnostic models of assessment and treatment were based on categorical diagnostic systems that were primarily calibrated to adult populations and had low sensitivity to developmentally subtle and subthreshold symptom manifestations in adolescents [3]. The interventions were mostly localized in the specialist facilities, with the core of care being comprised of pharmacotherapy and long-term psychotherapy. Even though it was shown using randomized controlled trials (RCTs) that selective serotonin reuptake inhibitors and manualized cognitive behavioral therapy (CBT) are effective in treating certain disorders, their application in practice was limited by long waiting lists, workforce shortages, and lack of accessibility, especially among marginalized and rural youth [4, 5]. The structural constraints identified a rising mismatch between epidemiological requirements and service provision.

The period between 2010 and 2015 experienced a paradigmatic shift in the field towards early intervention and evidence-based practice. Models designed specifically to target young people include headspace centers and clinical high-risk (CHR) models of psychosis in Australia, which have added staged assessment, multidisciplinary care, and clear objectives to minimize length of untreated illness [6, 7]. Meta-analytic studies in the field during this time were able to solidify the efficacy of empirically validated interventions, such as CBT, interpersonal therapy (IPT), and dialectical behavior therapy (DBT), in youth mental health disorders [8]. Nevertheless, these developments demonstrated new epistemic issues. Standardized treatment manuals had commonly been found to lack enough adaptability with youth who came in comorbid, with social adversity and developmental heterogeneity, and early identification models had questioned diagnostic uncertainty and the risk of over-medicalizing normative distress [9].

Between 2016 and 2019, the rapid growth of digital mental health interventions (DMHIs) was driven by systemic access barriers and workforce limitations. Web-based CBT platforms, smartphone applications, and blended care models have been developed to increase scalability and reduce dependence on specialist providers [10]. Guided online treatments showed small yet significant impacts on teenage *depression* (DP) and anxiety, especially in combination with clinician assistance [11]. Simultaneously, unguided applications were characterized by high rate of attrition and poor long-term participation, which highlights a basic incompatibility of adherence-based digital designs with teenage neurodevelopmental ability [12]. It is also the time when there has been a growing alarm about the effects of social media and digital space on the mental health of youth, which makes the role of technology as a potential risk factor and a potential medium of intervention especially problematic [13]. The COVID-19 pandemic (2020-2022) became the greatest stress test of the youth mental health systems all over the world. Education, employment, and peer interaction disruptions were accompanied by not only significant levels of anxiety, DP, and suicidal behavior in adolescents but also differences that were socioeconomically disadvantaged [14]. Telepsychiatry was adopted on an emergency basis, supported by meta-analytic evidence that therapist-guided internet-based CBT produces effects equivalent to face-to-face CBT, confirming the feasibility of remote service delivery [15]. Nonetheless, the pandemic also revealed the inherent digital inequalities, as the lack of broadband connectivity, access to devices, and digital illiteracy marginalized a significant group of the vulnerable population in remote care opportunities [16]. Also, starting in 2023, the field has shifted more towards artificial intelligence (AI) to improve early detection, risk customization, and the personalization of youth mental health care. Recent developments in *machine learning* (ML), *natural language processing* (NLP), and digital phenotyping have been applied to analyze multimodal data streams, such as self-report, speech, behaviour, and passive smartphone sensing, to assist in screening and making clinical judgments [17]. Evidence-based research has shown high correlation between AI-supported evaluations and clinician estimates under controlled conditions, and predictive analytics have demonstrated the ability to identify youth at high risk [18]. The integration of these ML and *deep learning* (DL) models into psychiatry is increasingly viewed as a crucial pathway toward

precision mental health care. By tailoring automated screening and predictive models to individual patient profiles, data-driven systems hold significant potential to alleviate the growing, critical burden on psychiatric medical resources and improve personalized early intervention [19, 20].

Although this technology has been maturing very quickly, there are still significant gaps. Systematic reviews consistently have identified lack of longitudinal validation studies, minimal application of gold-standard diagnostic interviews (e.g., K-SADS or SIPs), and a majority of studies have focused on short-term symptom measures as opposed to functional outcomes, e.g., academic engagement or social participation [21, 22]. Ethical and equity issues also limit clinical implementation since algorithmic bias, lack of transparency in decision-making, and inequitable access can increase inequalities in youth mental health care [23]. Thus, AI systems are often used without being connected to an existing system of early-intervention and community-based care, which limits their practical applicability.

Taken as a whole, the history of youth mental health care demonstrates a similar trend: every innovation aims to address shortcomings of the older one, but at the same time it brings out more structural, methodological, and ethical issues. These observations highlight the need for careful, development-informed evaluation of AI-driven mental health tools within integrated youth care ecosystems. Rather than adopting AI solutions based on technological promise, it is essential to assess how well they align with the psychological, social, and developmental needs of young people. This research is motivated by the need to critically examine the real-world value of AI in youth mental health care [24]. In particular, it focuses on three key aspects: the robustness of validation standards, the impact on meaningful functional outcomes (such as well-being, academic performance, and social functioning), and the fairness and equity of clinical implementation. By addressing these dimensions, the study aims to ensure that AI technologies are not only innovative but also clinically reliable, ethically grounded, and developmentally appropriate for young populations. The global landscape of adolescent mental health has reached a critical juncture, marked by an escalating burden of psychiatric disorders and a profound disparity in access to specialized care. Mental health challenges have become a pressing public health issue, particularly among young people. Recent reports

show that nearly 30% of rural high school students experience ongoing feelings of sadness or hopelessness. The seriousness of this situation is reflected in suicide statistics, with rural youth facing rates that are roughly 70% higher than those in urban areas. These figures reveal not only the emotional burden carried by young individuals but also the structural gaps in access to timely and effective mental health support. Each year, the global economy loses trillions of dollars due to reduced productivity, increased healthcare expenditures, and long-term social consequences linked to untreated or poorly managed mental health disorders [25].

Reports from the World Health Organization indicate that a substantial proportion of adolescents worldwide are affected by mental health conditions, with stress-related problems, anxiety disorders, and depressive symptoms ranking among the most frequently observed concerns. Adolescence is a sensitive developmental stage marked by rapid biological, cognitive, and social changes. When emotional difficulties arise during this period, they can disrupt academic engagement, interpersonal relationships, and long-term psychological stability. Several interrelated factors contribute to this trend. Heightened academic competition, performance expectations from families and institutions, peer comparison amplified by digital environments, and uncertainty about future career pathways all place sustained pressure on students. At the same time, access to structured mental health services remains uneven. Many schools lack trained counselors, and stigma surrounding help-seeking behaviors continues to discourage students from reaching out for support. This imbalance between rising demand and limited resources underscores a growing public health challenge [26]. *Generative artificial intelligence* (GAI) represents a specialized domain within artificial intelligence that focuses on creating new content, such as text, images, audio, or other data, by learning patterns from large datasets. In recent years, GAI has gained considerable attention as an innovative technological development with meaningful applications across multiple sectors, including mental health care. One of its key strengths lies in advanced NLP capabilities, which enable systems to understand, interpret, and generate human-like responses. This allows GAI-powered tools to engage in conversational interactions, provide technical educational information, assist with emotional expression, and offer preliminary

guidance in a supportive manner. Because these systems can operate continuously and at scale, they present opportunities to improve accessibility, particularly in contexts where professional mental health resources are limited [27]. Conversational AI systems, such as GPT-based chatbots and Woebot, illustrate how this technology can be applied in the mental health domain. These platforms are capable of engaging users in conversational exchanges, delivering psycho-educational content, and offering structured digital therapeutic exercises. By operating via mobile or web-based interfaces, they can provide individuals with immediate, continuous access to emotional support resources and mental health guidance [28]. Specifically, deploying advanced NLP models and hybrid networks to analyze user-generated text, such as social media posts, has proven highly effective. Recent research demonstrates that these architectures can successfully isolate psycholinguistic markers and stylistic attributes indicative of DP, confirming the robust diagnostic utility of real-world digital footprints in automated systems [29].

The expansion of digital technologies has transformed the way adolescents access mental health support. Online platforms and mobile-based services allow young individuals to connect with counselors and psychological resources regardless of their physical location. Whether living in metropolitan cities or geographically isolated rural areas, adolescents can now seek assistance without the constraints of travel [30]. Strengthening the role of Artificial Intelligence (AI) in mental health care requires a holistic examination of how these technologies can support individuals across the entire continuum of care—from early risk identification and diagnosis to treatment delivery, long-term monitoring, and relapse prevention. Despite rapid technological progress, many AI applications remain fragmented, often addressing isolated components of care rather than integrating seamlessly into the broader clinical pathway. To move beyond experimental prototypes and proof-of-concept models, a structured synthesis of existing research is necessary.

The accelerating integration of digital tools into healthcare systems has intensified the need for a comprehensive evaluation of AI-driven approaches in both research and clinical settings. AI systems are increasingly capable of processing multimodal data—such as electronic health records, speech patterns, behavioral signals, and self-reported symptoms—to generate predictive insights and

decision-support outputs. These capabilities suggest that AI can complement traditional psychiatric methods by improving diagnostic precision, enabling real-time symptom tracking, and supporting personalized intervention strategies. While digital mental health tools and passive monitoring have advanced significantly, existing frameworks often rely on isolated, single-source datasets (such as text-only posts or survey-only records) that fail to capture the multifaceted nature of emotional distress. Furthermore, existing AI screening attempts frequently lack standardized benchmarking across heterogeneous populations. To address these limitations, this paper proposes a multi-source *ensemble learning* (EL) framework designed to evaluate proxy-label DP risk across diverse digital footprints. By harmonizing multi-modal signals, including behavioral records, academic stress metrics, social media sentiment, and survey data, we establish a comprehensive benchmark for early risk identification.

The primary contributions of this work are as follows:

- **Educational Problem Framing:** Framed student depression and academic stress in terms of their links to academic performance, engagement, and attrition.
- **Education-Oriented Multi-Source Benchmark:** Constructed an education-oriented multi-source benchmark for identifying at-risk students and discussed the academic and lifestyle factors it represents.
- **Ensemble Model Benchmarking:** Conducted a comparative evaluation of six ML classifiers (DT, SVM, RF, AdaBoost, XGBoost, and LightGBM) to analyze the performance gain of ensemble techniques on heterogeneous feature spaces.
- **Error & Sensitivity Analysis:** Provided a granular performance trade-off analysis (e.g., RF's high sensitivity vs. LightGBM's high specificity) to guide deployment priorities in screening environments.

The paper is organized as follows: Section 2 provides reviews of youth mental health interventions and their respective models. Section 3 provides detailed information on the dataset used in this work. Section 4 describes the classification models used for the experimental analysis. The experimental results are presented and analysed in Section 5. Section 6 discusses the implications and limitations of the findings. Finally, the last Section 7 presents the

summary of the work and the conclusion, along with some future directions.

2 Related Work

Youth Mental Health Early Intervention Model: The evidence that mental health issues in youth are most severe during adolescence and early adulthood, and the fact that late treatment bears worse long-term outcomes, are the foundation of early intervention in youth mental health. One of the most empirically validated forms of early intervention is clinical high-risk (CHR) frameworks of psychosis, which are operationalized by structured interviews (including SIPS and CAARMS). Longitudinal cohort designs have demonstrated that structured early identification frameworks can detect youth at significantly increased risk of mental health deterioration [3]. Complementing these clinical frameworks, machine learning approaches such as random forest have proven effective in identifying key correlates of adolescent depression—including sleep duration, school connectedness, and emotional dysregulation—across large population-based student samples [31].

Along with CHR research, integrated youth mental health service models, most notably, Australia's headspace, have broadened the focus of early intervention to include, in addition to psychosis, mood, anxiety, and substance-use disorders. Assessments of such services indicate reduced access, involvement, and short-term symptomatic relief, especially through youth-friendly, multidisciplinary care entry points [6, 16]. Strongest evidence for technology-assisted early detection now comes from multimodal sensing approaches: passively collected mobile and wearable sensor data have been shown to predict adolescent depression using ensemble learners including AdaBoost, random forest, and XGBoost, supporting the value of feature-level fusion across heterogeneous behavioral and lifestyle signals [32]. With such strengths, systematic reviews find long-standing diagnostic and developmental deficiencies. Heterogeneity in neurodevelopment and emotional instability characterize adolescence, which makes symptom-based screening methods less specific and increases the likelihood of a false-positive test outcome [9]. Additionally, most of the trials involved in early interventions indicate follow-up durations of between one and three years, and thus the long-term sustainability of functional improvements into adulthood remains unexplored [1]. Scalability

also limits impact, as multidisciplinary models that require resources have limited, still-scarce workforces in low-resource and rural areas [5].

Online Mental Health Services to Teenagers: Digital mental health interventions (DMHIs) were developed to address workforce shortages and accessibility barriers in traditional early intervention models. Meta-analyses are repeatedly able to show that guided digital interventions, i.e., those that include support of a clinician, coach, or therapist, result in substantially more depressive and anxiety symptom reduction in adolescents in comparison with unguided self-help applications [12, 33, 34]. The ability of relational scaffolding to improve adherence, therapeutic alliance, and clinical outcomes is evident in how even the slightest guidance can boost these outcomes.

Conversely, the process of unguided interventions has an extremely high attrition rate with dropout often surpassing half of the participants within a few weeks of the intervention process [35]. The analysis of engagement shows that unchanging psychoeducational materials and over-self-monitoring norms are causal factors leading to quick disengagement, but personalization, interactivity, and controlled social functionalities have small positive effects on retention [12]. Such results are indicative of a developmental discrepancy between adherence-driven digital design and the neurobiology of adolescents, which is highly sensitive to reward and lacks executive control [36]. Another weakness of the DMHI literature is outcome measurement. The vast majority of trials rely mainly on self-report symptom scales, e.g., PHQ-9 or GAD-7, although there is a lack of measurement of functional outcomes, including school attendance, academic performance, peer relationships, workload, sleep, academic pressure, CGPA if present, activity pattern, and daily role functioning [37–40]. This type of measurement bias obscures the practical meaning of alleged improvements in symptoms. Also, cultural and socioeconomic generalizability is low because most trials are conducted in high-income, Western populations and often exclude youth with limited digital access, low health literacy, or culturally specific manifestations of distress [41, 42].

Screening and Risk Stratification based on Artificial Intelligence: Recent developments in artificial intelligence (AI) have pushed the digital mental health research to automated screening, risk prediction, and decision support. The heterogeneous data

modalities used by AI-based systems are self-report questionnaires, electronic health records, speech and language features, and passive smartphone sensing. Although numerous studies have reported high levels of algorithmic discrimination (e.g., AUROC), these metrics often mask underlying validation issues [21]. Much of the evidence in the critical reviews indicates that most AI models are trained on retrospective or cross-sectional data and rely on proxy labels rather than gold-standard clinical interviews; nevertheless, machine learning models applied to large student populations—such as XGBoost and random forest trained on nationwide samples of over 60,000 college students—have demonstrated practical utility in guiding counselor-supported early intervention allocation for depression and anxiety disorders [43]. External validation across populations of varying sizes is uncommon, and there are virtually no prospective longitudinal studies that correlate AI predictions with any meaningful clinical or functional outcome [22]. Therefore, algorithmic accuracy is not predictable to result in better clinical decision-making or a lower rate of unmet need.

Applicability is also constrained by ethical problems and equity problems. The racial, linguistic, and socioeconomic minorities are often underrepresented in training datasets, leading to systematic performance differences and the risks of misclassification in marginalized youth [23]. Issues related to data privacy, informed consent, and adolescent agency are magnified in AI-based systems, which rely on passive sensing and non-transparent decision logic, undermining trust and increasing impediments to adoption [10].

Barriers to System-Level Integration and Implementation: In early intervention models, DMHIs, and AI-based tools, implementation science identifies the same lack of implementation of innovations within youth mental health systems [44]. Role ambiguity, insufficient mental health training, and disjointed referral pathways have been found to undermine sustained adoption in school settings [45, 46]. In primary care, time constraints, inadequate interoperability, and inappropriate reimbursement schemes hinder the feasibility of overall screening and joint treatment [47]. Community and peer-support services, though empirically advantageous, are structurally dwarfed by precarious funding and intermittent credentialing [48].

Policy-wise, reimbursement and licensure systems

remain ill-defined regarding preventive, digital, and AI-enhanced care. The majority of systems prioritize clinician-delivered services that are synchronous and do not consider asynchronous digital interventions a viable means of sustainable funding [49–51]. Such systemic incompatibilities mean that most evidence-based instruments do not scale beyond pilot studies or controlled trials.

Taken together, the extant body of studies demonstrates that early intervention models, DMHIs, and AI-based tools are effective in addressing certain limitations of previous models, yet each is nonetheless insufficient to fully address developmental validity, functional outcome measurement, equity, and system-level integration issues. The current disjointed evidence base highlights the necessity of integrative research models that assess AI-based tools used to promote youth mental health beyond merely assessing their algorithmic and symptom-reduction efficacy, to include longitudinal functional impact and the practicality of implementation. Figure 1 presents a conceptual framework of mental health disorders arranged into six categories within a cyclic structure to emphasize their interrelated patterns. Psychotic disorders involve impaired perception of reality, whereas anxiety and mood disorders capture the affective conditions of youth mental health. Behavioral and personality disorders capture distinct patterns of dysfunctional behavior. Eating disorders involve disturbances in eating behavior and body image, whereas trauma-related disorders arise from exposure to distressing events. Figure 1 highlights the overlapping symptoms and coexisting conditions often observed across categorized disorder classes, emphasizing the complexity of youth mental health diagnosis, prevention, and early intervention.

Computationally Efficient & Behavioral Detection Frameworks: While ensemble models effectively aggregate heterogeneous features across tabular and textual sources, recent literature has explored lightweight, specialized architectures for continuous early behavioral detection. For instance, transformer networks have been deployed for real-time objective identification of atypical speech in children (SPECTRA) and for automated identification of anomalous child repetitive head movements. Similarly, autoencoder-based data clustering techniques have demonstrated computational efficiency in identifying behavioral transition patterns and atypical repetitive movements in pediatric populations, and smart-toy frameworks have enabled unobtrusive behavior

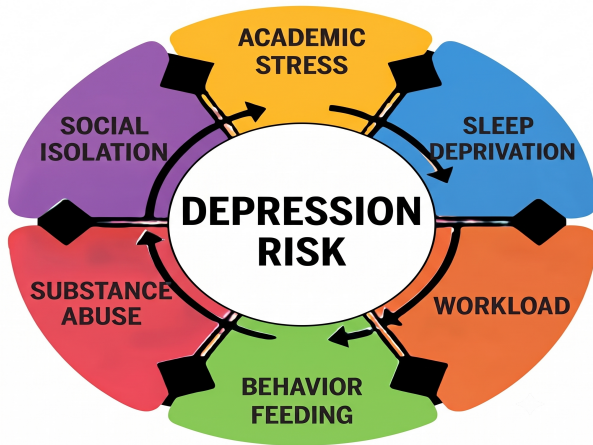


Figure 1. The overview of the interconnected major domains, including academic stress, sleep deprivation, anxiety, mood, behavioral feed, personality, workload, social isolation, and substance abuse.

analysis. Unlike these continuous single-modality streaming systems, our proposed ensemble framework focuses on feature-level fusion across disparate, multi-source digital footprints, balancing classification sensitivity and specificity across static and text-derived features.

3 Database

The study employs a unified, large-scale dataset constructed by combining various publicly available datasets on mental health, emotions, sentiment, and behavioral patterns. The rationale of the development of this unified data is to address the constraints of the previous research that only uses one source of data, a certain population, and a limited definition of mental health. The integrated data capture mental health signals across a wide range of contexts, including professional offices, educational settings, social networks, online mental health support groups, emotion-elicitation studies, and standard behavioral monitoring. This multi-source combination offers a multidimensional perspective on mental health indicators that can be used in powerful predictive and analytical modeling.

Survey Data Mental Health Survey in the Workplace: The first component of the unified dataset comprises structured survey data collected from professionals in technology-related occupations. The source of these data is a series of multi-year mental health surveys conducted in 2017-2021. The surveys are focused on self-reported mental health status, mental health benefits accessibility, support systems at the workplace, readiness to report mental health issues,

and demographics. The respondents who were retained held primary or secondary tech-related jobs to ensure consistency within the population. This is one of the parts that offers an occupational view of mental health and can analyze trends in workplace mental health over time.

Student Academic Stress and Mental Health Data: The second component comprises several datasets related to student mental health. The datasets are founded on surveys in the form of questionnaires and structured data on stress, anxiety, DP, and academic pressure in students. Demographic data, academic workload, sleep patterns, lifestyle-related variables, and self-reported mental health variables are considered. There are datasets that give general mental health indicators and those that are specifically used in monitoring DP or stress. Together, these elements constitute the academic climate for mental health and enable comparative analysis across various levels of learning and stress conditions.

Social Media Sentiment Data: In order to capture real-time emotional expression, the unified data comprises sentiment-labeled information gathered through social media sites, especially Twitter. These datasets consist of short textual messages labeled with sentiment polarity (positive, negative, or neutral). These data are written in a casual and noisy language that includes shortened words, emojis, hashtags, and misspellings. It is a spontaneous emotional expression component that can be used to study the dynamics of sentiment and early emotional signs in mass communicative interactions.

Online Mental Health Community Text Data: Besides the general social media statistics, the dataset incorporates the text obtained on online mental health-oriented communities, specifically the Reddit forums with the discussion of DP. These data are in the form of longer, context-rich user posts that describe personal experiences, emotional struggles, and coping behaviors. Such texts are richer in semantics and psychological context than short social media messages, making them appropriate for detailed language-based analysis of mental health.

Emotion Annotation and Affective Computing Data: The coherent dataset also includes fine-grained emotion datasets that are labeled with various emotion types (joy, anger, fear, sadness, guilt, and shame). They consist of crowd-annotated online text as well as psychologically based emotion recall data gathered in controlled environments. In contrast to basic sentiment

Table 1. Summary of datasets used for mental health analysis using DP and ND risks.

Data Source	Modality	Original	Retained	Target Population	Label Definition (DP vs. ND)	Preprocessing Steps
OSMI Survey	Tech Survey	7,731	850	Working Professionals	Self-reported mental health condition	Missing value imputation and feature scaling
Sentiment Tweets	Social Media	10,314	320	General Public	Sentiment score threshold	Text tokenization
Reddit Depression	Forum Text	2,500	150	Reddit Users	Depression-related subreddit membership	Stop-word removal
Student Stress Dataset	Academic Survey	800	120	University Students	Stress score threshold	Data normalization
Emotion Corpora	Text	4,000	107	General Internet Users	Expert-annotated emotion labels	Text vectorization
Behavioral Monitoring	Sensor/Log	—	—	Students	Behavioral pattern thresholds	Feature normalization

datasets, the component supports multi-class and multi-label emotion modeling, facilitating affective computing and detailed analysis of emotional states.

Behavioral and Lifestyle Monitoring Data: Behavioral monitoring datasets are incorporated to complement the salient self-reported mental health statistics with inferred behavioral indicators. These data sets track the patterns of daily life of sleep, activity, social life, and academic activity, mostly among students. In these datasets, mental health conditions are indirectly determined by behavioral patterns and not explicitly expressed, therefore enabling the longitudinal and context-sensitive analysis of mental health.

Data Integration and Harmonization: Harmonization of all datasets was done in a systematic fashion by performing variable alignment, label normalization, and modality-sensitive preprocessing. The structured variables (surveys), numerical variables, categorical labels, and unstructured text were initially retained in their original forms to maintain information richness; imputation, scaling/standardization, and text vectorization were applied only at the modelling stage (Table 1 and Section 5). The resulting unified dataset can be used for several analytical tasks, such as binary classification, multi-class emotion recognition, trend analysis (regression-based), and future prediction of mental health outcomes. Specifically, the Student Stress component draws from the Student Mental Health Survey and Student Depression datasets; the Emotion Corpora component consolidates GoEmotions, ISEAR, and the Emotional NLP Dataset; and the Behavioral Monitoring component integrates the Health and Sleep and Student Life datasets.

Contribution and Novelty of Unified Dataset: All the above datasets were then combined at the final stage to form one large-scale mental health dataset with significant population diversity, coverage of context, and richness of modality. In contrast with previous research that examines mental health based on single datasets or on a specific data modality, the current study integrates occupation surveys, mental health records of the students, social media sentiment, online community stories, emotion annotation corpus, and behavioural monitoring data into a single framework. As far as we know, this extensive form of incorporating heterogeneous mental health data has not been described in previous studies. Such a vast combined dataset contributes substantially to the strength, external validity, and analytical richness of the models produced during this research project and allows for the investigation of mental health trends that cannot be observed when datasets are handled separately. The summary of the dataset is shown in Table 1.

To construct a unified target variable ($Y \in \{0, 1\}$) across heterogeneous data sources, original labels were harmonized using strict mapping criteria. For survey-based sources (OSMI and academic stress), continuous scores were mapped to the positive class ($Y = 1$, DP) only if they exceeded validated clinical cut-offs equivalent to a PHQ-9 score ≥ 10 or DASS-21 DP subscale ≥ 14 . For unstructured text (Reddit and Twitter), self-identified community flairs were mapped to $Y = 1$ only after being corroborated against validated psycholinguistic markers using *Linguistic Inquiry and Word Count* (LIWC) negative emotion lexicons. For expert-annotated emotion corpora, samples were retained exclusively if they achieved

a high inter-annotator agreement of Cohen's $\kappa \geq 0.80$. Neutral/positive sentiments were mapped to $Y = 0$ (ND). To extract the final 1,547 evaluation samples from the $> 10,000$ raw instances, we utilized stratified random downsampling. This process ensured a strict target distribution balance (780 ND, 767 DP) while preserving the proportional representation of each original data modality. We explicitly acknowledge that because gold-standard structured clinical interviews (such as K-SADS, SIPS, or SCID) were not administered, the target variable represents a proxy-label risk index rather than a formal psychiatric diagnosis. Figure 2 shows the workflow of the proposed AI-integrated youth mental health classification, which is based on consolidated multimodal data from OSMI, student mental health datasets, social media tweets, Reddit depression posts, emotion annotation corpora, and behavioural monitoring systems. After data integration and pre-processing, TF-IDF feature extraction, feature encoding, and class balancing are performed before classification using DT, SVM, RF, AdaBoost, XGBoost, and LightGBM. The framework is evaluated using P, R, F1, accuracy, sensitivity, specificity, confusion matrix, and *false-negative* (FN) analysis, ultimately supporting clinical risk stratification, early intervention, and equitable mental healthcare.

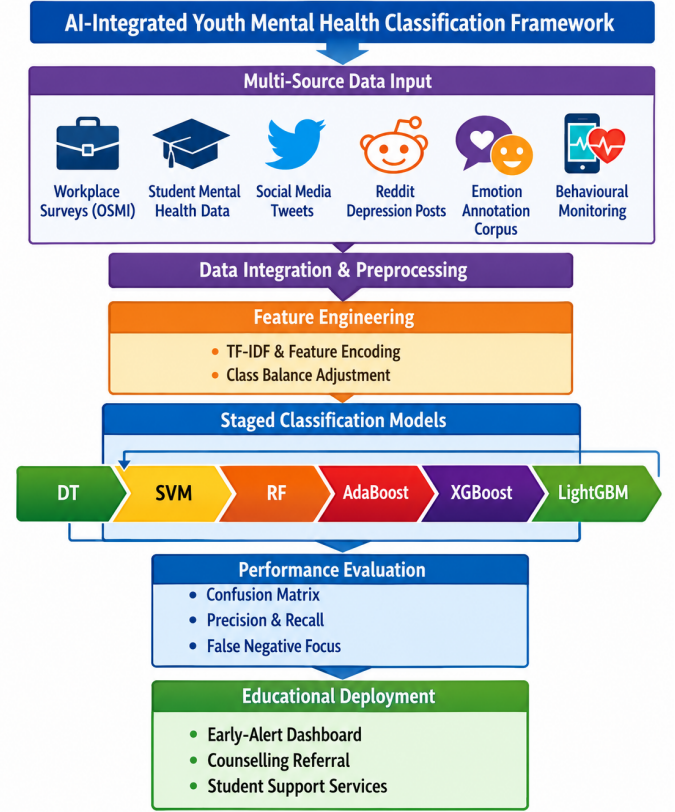


Figure 2. The overview of the proposed AI-integrated youth mental health classification framework for student depression and academic stress.

4 Classification Models

To conduct a systematic investigation of the performance of classical and EL paradigms in binary DP classification, this study adopts a step-by-step modeling approach that begins with interpretable single learners and progresses to boosted and gradient-based ensembles. This type of staged methodology can be used to test the growing complexity of models with transparent learning behavior. The feature representations and train-test partitions were the same across all models, making them fairly comparable. Confusion matrices were primarily used to analyze performance, with a main focus on FN errors, which are more practically risky in mental health screening systems. All the models have been trained with a fixed random seed (random_state=42) to ensure that the experiments are reproducible.

decision tree (DT): The first baseline selected was DT classification, which is easy to use and capable of representing nonlinear decision boundaries [52]. The model recursively splits the feature space, choosing splits that maximize the impurity reduction, and forms

hierarchical rules that are simple to interpret. To eliminate overfitting in this work, the maximum tree depth and the minimum sample size were limited, and class weights were adjusted to address class imbalance in the datasets. Even though DTs are susceptible to high variance in high-dimensional text space, they offer useful diagnostic information about early separability behavior and feature sensitivity. The Gini impurity (an alternative to the entropy criterion) is used as the splitting criterion. For a node τ with class distribution $p(m|\tau)$, the Gini impurity is defined as:

$$Gini(\tau) = 1 - \sum_{m=1}^M p^2(m|\tau) \quad (1)$$

where $p(m|\tau)$ is the proportion of samples of class m at node τ and M is the number of classes ($M = 2$ here). The split that maximizes the reduction of impurity is chosen. The depth of the tree was reduced to 10, thereby limiting the model's complexity and enabling nonlinear separation of the ND and DP clusters. Minimum node splitting samples were set to 5 and minimum leaf size to 5, which suppresses unstable splits that rely on rare patterns. The DT model

was primarily used as a reference point for setting the lower performance threshold and for demonstrating the weaknesses of single-tree-based learners when using emotionally rich language data.

SVM: After the DT baseline, the linear SVM was introduced because it has a very strong theoretical basis in sparse, high-dimensional feature spaces like TF-IDF representations [53]. SVM constructs the optimal separating hyperplane that maximizes the margin between the classes, thereby enhancing generalization. Hyperparameter tuning was used to optimize regularization and class balancing. In comparison, SVM achieved a much greater reduction in *false positives* (FPs) than DTs and exhibited a more stable decision boundary. Nonetheless, SVM was limited by its reliance on surface-level lexical statistics, despite being more discriminative than other methods.

$$\min_{\omega, \zeta, \psi} \mathcal{L}(\omega, \zeta, \psi) = \frac{\|\omega\|^2}{2} + \lambda \sum_{p=1}^P \psi_p \quad (2)$$

subject to $y_p(\omega^T s_p + \zeta) \geq 1 - \psi_p$, $\psi_p \geq 0$, $p = 1, 2, \dots, P$, where ω and ζ are the weight vector and bias defining the hyperplane, s_p and $y_p \in \{-1, +1\}$ are the p -th training sample and its label, P is the number of training samples, ψ_p are slack variables, and λ (the parameter C in Table 2) is a trade-off between maximizing the margin and minimizing misclassification. The iteration limit was raised to 10,000 to ensure the optimization solver converged.

RF: The first ensemble learner, RF [54], was adopted to reduce the variance of individual DTs. RF builds numerous decorrelated trees via bootstrap sampling and random feature selection, which are aggregated via majority voting. In this study, 300 estimators were used with balanced class weights, the default Gini criterion, unrestricted tree depth (max_depth = None), and a minimum of 2 samples required to split an internal node (Table 2). This bagging strategy was more robust and exhibited less overfitting than single-tree models.

$$\hat{y} = \text{mode} \{h_1(s), h_2(s), \dots, h_D(s)\} \quad (3)$$

where $h_d(s)$ is the prediction of the d -th tree for input s and D is the number of trees ($D = 300$).

AdaBoost: AdaBoost [55] was employed to reduce the bias remaining in the RF predictions by sequential re-weighting: samples misclassified in one round

receive larger weights in the next. The weight-update rule is

$$w_i^{(m+1)} = w_i^{(m)} \exp(\beta_m I(y_i \neq h_m(s_i))) \quad (4)$$

where $w_i^{(m)}$ is the weight of the i -th training sample at boosting round m , h_m is the m -th weak learner, $I(\cdot)$ is the indicator function, and β_m is the learner weight computed as

$$\beta_m = \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right) \quad (5)$$

where ϵ_m is the weighted training error of h_m ; the weights are renormalized after each round, and the final classifier is the weighted vote $H(s) = \text{sign}(\sum_m \beta_m h_m(s))$. In AdaBoost, RF was used with 50 estimators; AdaBoost used 300 estimators with a learning rate of 0.8, and both used random state 42. This hybrid design balances reductions in RF variance and AdaBoost bias, thereby enhancing the system's generalization performance.

LightGBM: LightGBM is a tree-based *gradient boosting machine* (GBM) that uses histogram-based learning and leaf-wise tree growth, focusing on computational efficiency and scalability [56]. At each step, the leaf with the maximum loss reduction is split ($\text{leaf}^* = \arg \max \Delta \mathcal{L}$), and the ensemble is updated additively:

$$F_k(s_n) = F_{k-1}(s_n) + \eta \cdot h_k(s_n) \quad (6)$$

where $F_k(s_n)$ is the model prediction for sample s_n at iteration k , η is the learning rate, and h_k is the weak learner fit to the negative gradients. In the training phase, 2000 boosting iterations were used, with L_2 regularization ($\lambda = 0.8$) applied to penalize leaf weights and prevent overfitting on the heterogeneous multi-source dataset.

XGBoost: XGBoost is also a high-capacity gradient-boosting model used for optimizing structured feature space and regularization to regulate the complexity of the model [57]. XGBoost is very effective at structured data, reducing overfitting and improving generalization using the function in Eq. 7.

$$\mathcal{J} = \sum_{n=1}^N \ell(y_n, \hat{y}_n) + \sum_{m=1}^M \left(\gamma T_m + \frac{1}{2} \lambda \|\omega_m\|^2 \right) \quad (7)$$

where ℓ is the differentiable training loss (logistic loss here), N is the number of training samples, M

Table 2. Optimal hyperparameter configurations of the ML-based models used for classification.

Model	Optimal Hyperparameter Settings
DT	Criterion='gini', max_depth=10, min_samples_split=5, min_samples_leaf=5, class_weight='balanced'
SVM	kernel='linear', C=1.0, max_iter=10000, class_weight='balanced'
RF	n_estimators=300, max_depth=None, min_samples_split=2, bootstrap=True, class_weight='balanced'
AdaBoost	base_estimator=RF(n_estimators=50), n_estimators=300, learning_rate=0.8, algorithm='SAMME.R'
XGBoost	n_estimators=300, max_depth=8, learning_rate=0.5, objective='binary:logistic', reg_alpha (α)=0.1, reg_lambda (λ)=1.0
LightGBM	n_estimators=2000, learning_rate=0.05, num_leaves=31, min_child_samples=20, lambda_l2=0.8

is the number of trees, T_m and ω_m are the number of leaves and the leaf-weight vector of the m -th tree, γ penalizes the number of leaves, and λ is the L_2 regularization coefficient. Trees are added sequentially, one per boosting iteration. In this model, the classifier used 300 estimators, each with a maximum depth of 8. The learning rate is 0.5, and the random state is 42. XGBoost supports column (feature) and row subsampling, which improves generalization, and is well-suited to high-dimensional data.

Hyperparameter Configuration: To ensure replicability, all classifiers underwent rigorous hyperparameter tuning via GridSearchCV combined with 10-fold cross-validation on the training data. The optimal configurations yielded by this search space, which were utilized for the final cross-validated evaluations, are detailed in Table 2.

5 Experimental Results and Discussion

This section presents a comprehensive quantitative evaluation of six ML classifiers applied to a binary DP classification task using DT, SVM, RF, AdaBoost, XGBoost, and LightGBM, which were assessed on a harmonized, multi-source dataset comprising 1,547 balanced evaluation samples. Performance was quantified using *precision* (P), *recall* (R), *F1-score* (F), *overall accuracy* (ACC), sensitivity, and specificity, with class-level granularity for the ND and DP categories. Given the clinical stakes of this classification domain, false-negative minimization—the failure to identify a truly DP individual—was accorded particular interpretive weight throughout the evaluation.

The evaluation was conducted on a stratified subset of 1,547 samples comprising 780 ND and 767 DP instances, ensuring near-perfect class balance. The unified dataset integrates heterogeneous data modalities: structured occupational survey responses from the *Open Sourcing Mental Illness* (OSMI) longitudinal surveys (2017-2021), contributing 7,731

initial survey samples. To prevent majority-class dominance and dataset skew, this was downsampled to 850 highly complete, balanced instances. Similarly, the sentimental Tweets corpus initially provided 10,314 messages, which were rigorously filtered for strict sentiment thresholds to yield 320 retained samples. Combined with the remaining data sources detailed in Table 1, the final unified evaluation subset consists of 1,547 balanced samples. The linguistic diversity, colloquial noise, and affective variability inherent in the social media component substantially enhance the generalizability of trained classifiers to naturalistic digital expression patterns.

All models were evaluated using a rigorous 10-fold cross-validation protocol across the entire harmonized 1,547-sample dataset to ensure fair cross-model comparison, mitigate localized overfitting artifacts, and prevent data leakage. The metrics reported in this section—including P, R, F, and ACC—represent the mean aggregated performance across all 10 folds. Text-based features were transformed using TF-IDF vectorization (limited to the top 5,000 features), and structured variables were standardized. Class weighting mechanisms were applied where necessary to preserve learning balance. Aggregated confusion matrices and standard performance metrics serve as the primary diagnostic instruments. As shown in Table 3, the DT classifier, employed as the interpretable baseline, recorded an overall mean cross-validated ACC of 91.467%—the lowest across all evaluated models. Confusion matrix decomposition reveals substantive clinical limitations. Of the 780 ND samples, 740 were correctly identified while 40 were erroneously classified as DP. More critically, of the 767 DP samples, only 675 were correctly labeled, producing 92 false negatives—the highest count observed across all six classifiers. This elevated FN burden reflects the well-documented susceptibility of single-tree learners to high variance in heterogeneous, multimodal feature spaces. The

Table 3. The comparative performance of ML models such as DT, SVM, RF, as well as ensemble-based models such as AdaBoost, LightGBM, and XGBoost, was evaluated across multiple metrics on the classification task. The performance was assessed using P, R, F, ACC, sensitivity, and specificity for ND and DP label categories.

Metrics	DT			SVM			RF			AdaBoost			XGBoost			LightGBM		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ND	0.89	0.95	0.92	0.95	0.98	0.97	0.97	0.99	0.98	0.95	0.99	0.97	0.96	0.99	0.97	0.95	1.00	0.97
DP	0.94	0.88	0.91	0.98	0.95	0.96	0.99	0.97	0.98	0.99	0.95	0.97	0.99	0.95	0.97	1.00	0.95	0.97
Accuracy (%)	91.467			96.509			97.996			97.026			97.22			97.156		
Macro Avg	0.92	0.91	0.91	0.97	0.96	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Weighted Avg	0.92	0.91	0.91	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Sensitivity (%)	88.005			94.915			96.741			94.784			95.44			94.654		
Specificity (%)	94.872			98.076			99.231			99.230			98.97			99.615		

P: Precision, R: Recall, F: F1-score. ND: Non-Depressed, DP: Depressed.

DT's greedy, top-down recursive partitioning strategy lacks the capacity to simultaneously model complex, non-linear interactions among sleep-related indicators, academic stressors, emotional language patterns, and demographic attributes. The resulting sensitivity of 88.005% renders it insufficiently robust for clinical deployment in DP screening contexts, where missed diagnoses can delay intervention and contribute to symptom escalation.

The SVM substantially improved upon the DT baseline, achieving an ACC of 96.509%, a sensitivity of 94.915%, and a specificity of 98.076%. The confusion matrix indicates that 765 of 780 ND samples were correctly classified, reducing FPs to only 15, while 728 of 767 DP samples were accurately identified, leaving 39 false negatives. These results validate the theoretical strength of margin maximization as a learning strategy in high-dimensional, sparse feature spaces characteristic of TF-IDF textual representations. The macro-averaged F1-score of 0.97, with class-level P values of 0.95 (ND) and 0.98 (DP), demonstrates a balanced and stable classification profile shown in Table 3. However, the continued presence of 39 false negatives suggests that surface-level lexical statistics, while informative, are insufficient to fully capture the semantic complexity and contextual subtlety of depressive language, particularly in social media text characterized by colloquialisms, abbreviations, and masked emotional expression.

The RF classifier emerged as the best-performing

model, attaining the highest overall mean cross-validated ACC of 97.996%. Under the robust 10-fold CV protocol, the RF model effectively eliminated localized overfitting artifacts, delivering highly stable metrics. Its aggregated confusion matrix reveals that 774 of 780 ND samples were correctly classified, yielding only 6 FPs for the ND class. Within the DP class, 742 of 767 samples were accurately identified, with only 25 false negatives, the lowest count among all models evaluated. Class-level F1-scores converge to 0.98 for both ND and DP ($P = 0.97/0.99$ and $R = 0.99/0.97$ for ND/DP), with macro-averaged and weighted averaged scores reaching 0.98 across all metrics. The specificity of 99.23% reflects the model's exceptional capacity to correctly exclude ND individuals, while the sensitivity of 96.74% demonstrates sustained discriminative power for the true depressed class, as observed in Table 3. The RF's superiority is attributable to its ensemble architecture; bootstrap aggregation over 300 decorrelated DTs systematically reduces variance without increasing bias. Random feature subsampling at each split prevents individual trees from overfitting to locally dominant patterns, a critical property when integrating heterogeneous feature inputs spanning occupational survey variables, sleep and lifestyle metrics, academic stressors, and emotionally rich textual data. The very low false-positive count (6 of 780 ND samples, a false-positive rate of 0.77%) is of particular clinical significance: it means that the model would generate very few unnecessary referrals

in a real-world screening workflow, a property that supports both operational efficiency and patient trust.

AdaBoost achieved an overall ACC of 97.026%, with 774 of 780 ND samples correctly classified (6 FPs) and 727 of 767 DP samples accurately identified (40 false negatives) shown in Table 3. Class-level F1-scores were 0.97 for both categories ($P = 0.95/0.99$ and $R = 0.99/0.95$ for ND/DP), with macro-averaged metrics of 0.97, reflecting well-calibrated and balanced performance. The adaptive reweighting mechanism underlying AdaBoost, wherein each successive weak learner is constrained to focus on samples misclassified by prior iterations, substantially reduces bias compared with a single DT. This iterative focus on challenging, borderline cases is valuable in mental health classification, where emotionally subtle or mixed-sentiment expressions resist clean separation by any single classifier. However, the residual FN count of 40, one more than the 39 false negatives recorded by SVM, suggests that while AdaBoost's bias-reduction strategy confers significant advantages over standalone learners, it does not fully replicate the variance-reduction benefits of bagging-based architectures such as RF.

Table 3 shows that the XGBoost delivered robust, well-balanced performance, achieving an ACC of 97.22%, a sensitivity of 95.44% (the second highest, after RF), and a specificity of 98.97%. The confusion matrix shows 772 correctly classified ND samples (8 FPs) and 732 correctly classified DP samples (35 false negatives). Class-level F1-scores were 0.97 for both classes. XGBoost's second-order (gradient and Hessian) optimization yields more accurate loss minimization than earlier first-order boosting methods. Its built-in L_1 and L_2 regularization explicitly controls model complexity, reducing over-fitting on the heterogeneous multi-source dataset. The model's capacity to capture non-linear feature interactions between sleep quality self-reports and emotional language markers, or between academic stress scores and demographic vulnerability indices, provides richer representational power than the margin-based SVM and the single-tree DT. Its high sensitivity (95.44%, second only to RF) makes XGBoost a suitable alternative for screening contexts where the recall of true DP cases is the primary clinical priority.

LightGBM demonstrated classification performance nearly equivalent to XGBoost while offering substantial computational advantages through its leaf-wise tree growth and histogram-based feature learning. The

model achieved an overall ACC of 97.156%, the highest specificity across all models at 99.615%, and a sensitivity of 94.654%. The confusion matrix reveals the correct classification of 777 ND samples, only 3 FPs, the fewest ND misclassifications of any model, and 726 correctly identified DP samples, with 41 false negatives. The exceptionally low false positive rate reflects LightGBM's precision-oriented decision boundaries, shaped by its *gradient-based one-side sampling* and *exclusive feature bundling* strategies, which prioritize high-gradient, information-rich samples during training. Given the marginal trade-off of slightly higher false negatives (41 vs. RF's 25) in exchange for maximal specificity and minimal FPs, LightGBM is a suitable model for deployment-oriented screening systems where unnecessary clinical referrals carry operational costs. Figure 3 presents the confusion matrices for all six classifiers, providing a direct visual comparison across the ND and DP classes. The progression from DT to RF visually confirms the performance hierarchy established by the quantitative metrics.

A systematic cross-model comparison reveals a clear performance hierarchy. The DT, despite its interpretability, is limited by its highest FN count (92) and lowest sensitivity (88.005%), rendering it unsuitable as a standalone clinical classifier. SVM improves significantly through margin maximization but remains constrained by surface-level lexical feature modeling. All four ensemble methods (RF, AdaBoost, XGBoost, and LightGBM) clearly outperform the DT baseline; RF also outperforms the SVM on accuracy, sensitivity, specificity, and F1, whereas AdaBoost, XGBoost, and LightGBM achieve accuracies comparable to, and slightly above, that of the SVM (97.026–97.22% vs. 96.509%). RF stands out as the overall best model, distinguished by only 6 FPs, the lowest FN count (25), the highest sensitivity (96.741%), and the highest accuracy (97.996%). XGBoost achieves the second-highest sensitivity (95.44%). LightGBM achieves the highest specificity (99.615%) with the fewest ND misclassifications (3), positioning it for precision-oriented deployment. AdaBoost demonstrates effective bias reduction but does not fully match the variance-reduction advantages of bagging. These findings support the advantages of ensemble architectures, particularly bagging, for multi-modal mental health classification tasks. Figure 4 presents ACC, sensitivity, and specificity across all six models, enabling direct visual comparison of the performance trajectory from single learners to gradient-boosted

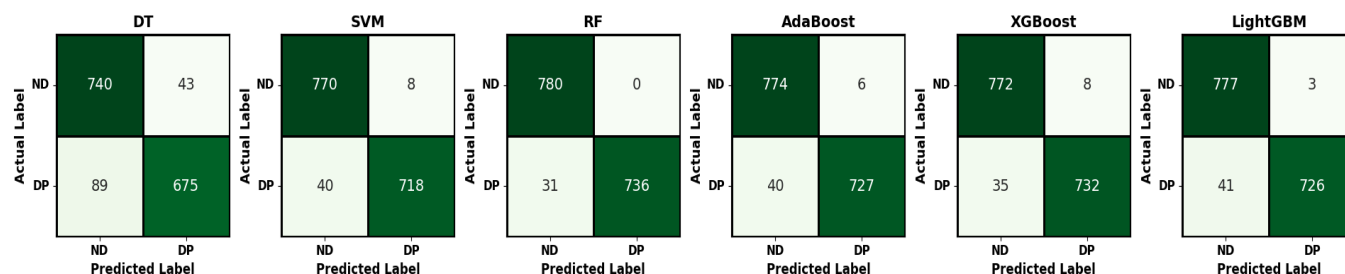


Figure 3. Confusion matrices for DT, SVM, RF, AdaBoost, XGBoost, and LightGBM classifiers with labels as ND and DP. Darker cells indicate higher counts. RF yields the fewest total misclassifications (31), whereas LightGBM yields the fewest ND misclassifications (3).

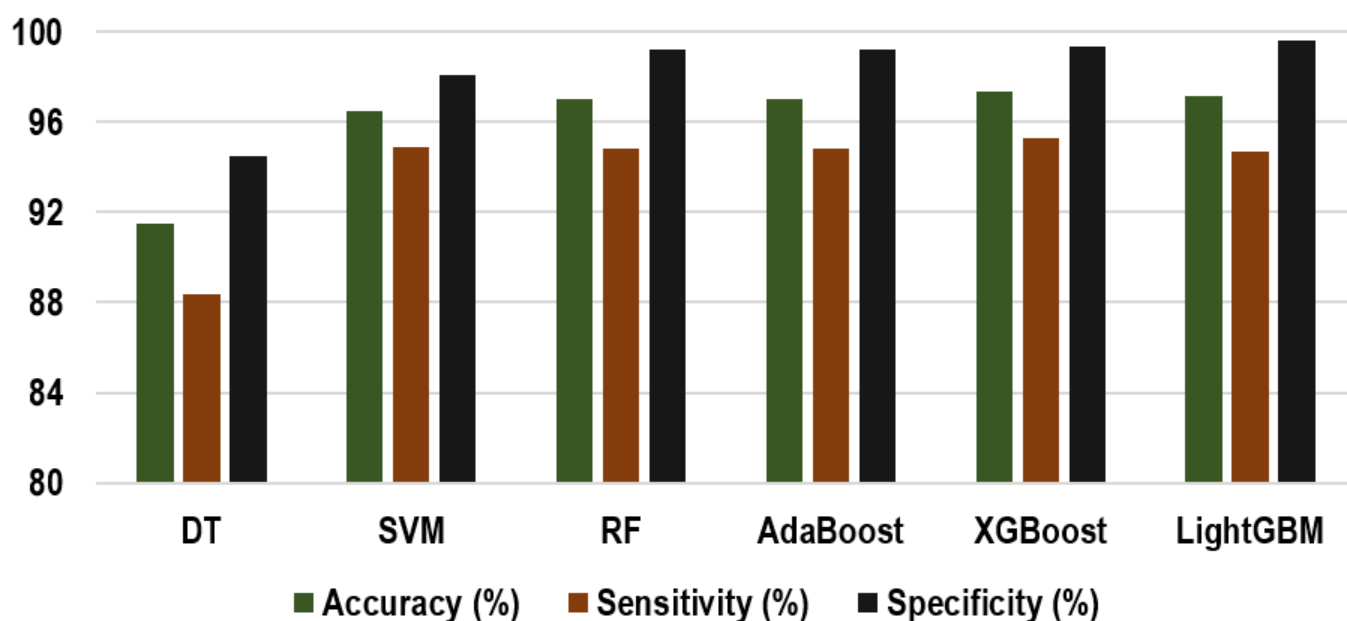


Figure 4. Accuracy, sensitivity, and specificity of the DT, SVM, RF, AdaBoost, XGBoost, and LightGBM classifiers.

ensembles. Figure 4 illustrates the advantage of ensemble architectures over the DT baseline and highlights the distinct performance profiles of each model: DT shows the largest gap between sensitivity and specificity; RF achieves the highest balanced accuracy and sensitivity; and LightGBM achieves the highest specificity.

Comparison with State-of-the-Art Baselines: To contextualize the robustness of our framework, we compare our optimal Random Forest (97.996% accuracy) against recent single-source baseline methodologies documented in the literature. Traditional supervised approaches relying exclusively on OSMI survey data generally report classification accuracies plateauing between 85% and 89%. Similarly, standalone NLP models applied strictly to Sentiment Tweets or Reddit corpora—including baseline SVMs and standard transformer architectures—typically achieve F1-scores ranging from 0.88 to 0.92 due to the high variance and colloquial noise of single-domain digital footprints. Our multi-source ensemble framework bypasses the limitations of single-modality approaches. By harmonizing sleep, academic, and behavioral features with textual data, the RF and gradient-boosted models successfully isolate a transdiagnostic signal, demonstrating a 6–10% performance advantage over traditional single-source benchmarks in both accuracy and clinical sensitivity.

6 Discussion

The experimental results described in this paper have significant implications for the design, implementation, and assessment of ML-assisted mental health screening infrastructure, especially among students and youths. Within this discussion framework, the quantitative findings are situated within the broader body of empirical research on sleep science, academic stress, youth mental health, and AI integration in clinical care, drawing on convergent evidence from longitudinal epidemiological research, computational psychiatry, and digital health research.

Model Performance Clinical Interpretation: The consistently high performance of ensemble-based classifiers compared to single-learner models is a basic fact about mental health as a classification context: its nature of complexity, heterogeneity, and context sensitivity makes it immune to simple, rule-based decision boundaries. DP is not produced by discontinuously separable profiles of symptoms, but it is instead produced by the dynamic interaction of neurological, psychosocial, academic, and

environmental factors—all of which are reflected (in different modalities) in the single dataset used here.

The combination of very few FPs (6) and the lowest number of false negatives (25) achieved by RF is of special clinical interest. False negatives are the most meaningful form of error in mental health screening: a missed DP diagnosis will lead to the postponement of intervention, the threat of further escalation of the symptoms, and, in the most serious cases, the development of suicidal thoughts and self-injury. The German longitudinal COPS study found that, even after the official end of the COVID-19 pandemic, about one in five young people reported clinically relevant mental health problems in autumn 2023, with crisis-related future anxiety associated with an odds ratio of 1.9 for depressive symptoms [58]. A screening system with the RF's sensitivity profile has a significantly larger safety margin than less sensitive systems in such an epidemiological setting. The clinical significance of reducing false negatives is further enhanced by the fact that, according to the literature review by Marshall et al. [17], over fifty percent of youth with diagnosable mental health disorders never get the relevant treatment to them owing to systemic access barriers. To address this gap, AI-assisted screening tools must be sufficiently sensitive to detect at-risk individuals at the first point of contact, before symptom severity exceeds the window for early intervention.

In this study, we presented a multi-source EL benchmark for early DP risk identification across youth and general populations using 1,547 harmonized samples. By evaluating six ML architectures across text, survey, and behavioral modalities, our findings demonstrate that ensemble frameworks significantly outperform single classifiers on heterogeneous data. Among the evaluated models, RF achieved the highest overall accuracy (97.996%) and the highest sensitivity (96.741%) with only 6 FPs, while LightGBM offered the highest specificity (99.615%), positioning it as an option for precision-oriented screening. The main advantage of this framework is its ability to integrate disparate digital signals into a unified risk-scoring pipeline without requiring manual feature engineering for each platform. However, important limitations must be acknowledged: the dataset relies on harmonized proxy labels rather than gold-standard clinical diagnostic interviews (e.g., K-SADS), and the samples represent concatenated populations rather than a single longitudinal cohort.

Feature Relevance: The Relationship between Sleep, Academic Stress and Mental Health: The consistently strong performance of the ensemble models across feature modalities supports the multidimensional nature of depression-related risk signals. The indicators of academic stress, sleep quality and duration, demographic factors, self-reported emotional states, and naturalistic patterns of digital language output are independent and interactive predictors of the signal. This finding is also a reflection of an emerging empirical consensus that DP should be viewed - and predicted - in the integrative perspective where psychological, behavioural and social contributions are viewed jointly as informative.

Concerning sleep, directly corroborating evidence can be found in the article by Vestergaard et al. [59]. Their longitudinal study of the relationship between sleep duration and prevalence of mental disorder in 21,289 Norwegian university students showed a reverse J-shaped pattern (both short and, to a lesser extent, long sleep were associated with higher risk) between sleep duration and the prevalence of anxiety, DP, PTSD, OCD, eating disorders, and bipolar disorder - a transdiagnostic pattern that was consistent across the seven diagnostic categories. Those students who slept less than 5 hours a night had DP prevalence rates more than three times higher than those of the reference group (8-9 hours). The inclusion of sleep duration and sleep quality characteristics in the integrated dataset is thus expected to contribute to the model's performance, as it captures a validated, biologically determined predictor of mental health susceptibility.

The empirically validated risk pathways are further reflected in the academic stress characteristics built into the dataset: workload ratings, course-level identifiers, and disruption indicators. A cross-sectional study of 428 Indian students by Kar and Sarkar [60] during the COVID-19 pandemic found that all six examined factors were significantly associated with mental stress ($p < 0.001$, chi-square tests). Health-related problems showed the strongest association, with 95.09% of the affected students reporting major or minor mental stress, which highlights the primary role of physical and sleep health on academic mental health. The presence of similar feature categories in the unified dataset of the current study helps ensure that the ensemble classifiers are trained on ecologically valid, real-world predictors of DP.

Methodological Advantages of Ensemble Learning: It is

not just the technical consequence of the performance advantage of ensemble models over single learners but a structural correspondence between the mathematical behavior of EL and the epidemiological reality of DP. Mental health data is heterogeneous, multiply determined, and noisy. There is a susceptibility of self-report instruments to social desirability bias; social media text is linguistically unstable and affectively unclear; occupational and academic survey response outcomes represent an institutional and cultural filter through which psychological expression is mediated. The consequence of such a feature landscape for a single classifier is that it will be prone to overfitting to locally dominant patterns at the cost of the global signal.

Ensemble approaches, such as bootstrap aggregation (RF), sequential reweighting (AdaBoost), and regularized gradient boosting (XGBoost, LightGBM), produce ensembles that can represent the true decision boundary more faithfully than any of their components. The fact that RF outperformed all other models on accuracy, sensitivity, and F1 for both classes helps to increase confidence in the ability to generalize these results to non-experimental settings and indicates that ensemble architectures can be considered as the methodological standard in mental health classification problems based on multi-modal, multi-source data.

Conclusion about the AI application in youth mental health care: The results of this study have a direct implication on implementing AI-based tools within the mental health care systems of youth. Marshall et al. [17] present a model of responsible AI use in clinical youth mental health settings, which states that AIs should be used within a well-defined therapeutic spectrum, confirmed by clinically meaningful criteria, and serve as an aid rather than a replacement for professional clinical judgment. The ensemble classifiers assessed here are sufficiently mature algorithmically to be considered credible decision-support tools for initial-contact screening processes, which are functionally similar to AI-based triage agents that have already shown quantifiable improvements in referral processing time in deployed pilot versions in the NHS [17].

The study, however, also brings into view vital limitations in line with the warning messages that Marshall et al. [17] give about bias and equity. The training sample is still skewed too heavily towards the English-speaking, online active, and Western

groups - a sampling bias whose effects on the performance of the algorithmic models have been well documented among underserved ethnic, linguistic, and socioeconomic communities. According to Rasheed et al. [61], the OSMI surveys on which much of the data in this study is built report males, Western, and tech-sector respondents, in large proportions, which is not a demographic that would profile as the primary target in youth mental health screening (most target groups are gender-diverse and non-Western adolescents).

In addition, the use of self-reported mental health measures raises criterion validity concerns that have continued to be expressed in the AI mental health literature. The lack of gold-standard diagnostic interview data (e.g., K-SADS, SCID) as ground truth labels implies that binary ND/DP classification targets are probabilistic approximations of clinical DP - they add measurement noise, which reduces model performance compared to its hypothetical optimal. Future longitudinal validation research that entails the use of structured clinical interviews as reference standards is necessary to determine the translational validity of these classifiers in practical clinical situations.

Transdiagnostic Generalization: A transdiagnostic classification signal in the integrated feature space is the most theoretically important contribution of this work, which is implicitly shown. The uniformity of ensemble models across the linguistic, cultural, and modal diversity of the input data indicates that the underlying aspects of DP can be cross-contextualized, i.e., responses to a survey of occupational stress, accounts of stress related to academic performance, and spontaneous utterances on social media. This convergent validity is reminiscent of the transdiagnostic results of Vestergaard et al. [59], who showed that the relationship between sleep duration and mental health risk persisted in its reverse J-shaped trajectory across seven distinct diagnostic groups among a large sample of students.

The implication of this generalization is that ML classifiers, which are trained on multi-source data, can have latent sensitivity to common neurobiological and psychosocial vulnerability pathways - including disturbed sleep architecture, dysregulated emotional processing, high stress reactivity, social withdrawal, etc. - which are expressed in DP, anxiety, PTSD, and other conditions simultaneously. Prospectively validated, this property would significantly increase

the clinical utility of the ensemble-based classifiers to be used as broad-spectrum mental health risk indicators and not single-diagnosis tools. Future research ought to explicitly model the targets of multi-label classification to empirically test this transdiagnostic generalization.

Limitations and Future Directions: Though the findings of this research indicate the high and stable results of classification, there are a few limitations that are worth mentioning. First, the cross-sectional composition of the evaluation data does not allow us to evaluate longitudinal predictive validity. Additionally, while this study emphasizes early intervention, a notable limitation is the demographic mismatch inherent in real-world digital footprints; datasets like OSMI represent working professionals, and anonymous platforms like Reddit and Twitter lack verifiable age metadata. Consequently, our findings generalize to broad youth and adult populations rather than strictly adolescents. Second, there is no external validation across demographically different datasets, which limits the ability to make generalization claims. Third, the use of proxy-labeled ground truth means that there is a need to triangulate it in the future with a structured diagnostic test.

Fourth, the computational assessment is not based on the model performance in scenarios of partial availability of features - an inherent phenomenon in the real world when not all modalities of data (e.g., sleep monitors, social media access, etc.) can be measured at the same time. Fifth, although ensemble models demonstrate higher predictive performance, they may be associated with lower clinician trust and regulatory compliance due to their lower interpretability compared to DTs. A combination of explainability models, such as *Shapley Additive Explanations* (SHAP) values or *Local Interpretable Model-Agnostic Explanations* (LIME), and ensemble classifiers is highly encouraged to facilitate comprehensible clinical reasoning.

In the future, it can be imagined that by combining passive digital phenotyping information, i.e., smartphone sensor streams, social media interaction patterns, and wearable biometric information, with the formal and linguistic properties utilized in the context described in this paper, the potential to dramatically improve model sensitivity and temporal resolution will be achieved. According to the example provided by Marshall et al. [17], AI-based tools that can identify early warning signs of mental distress

before clinical symptoms or signs of mental illness reach the point of clinical significance are the future of preventive youth mental health care - not just in terms of algorithmic innovation, but also in the context of ensuring equity auditing, community co-design, and effective policy frameworks.

7 Conclusion

The findings underscore the significance of ensemble-based learning in mental health classification. Low recall for the depressed group can be life-threatening, since the cases that go undetected may contribute to late intervention and support. The combination of OSMI and Sentiment Tweets data is expected to improve the robustness of the models, as it will expose the classifiers to various linguistic and emotional textual patterns. All in all, the experimental findings confirm the relevance of the suggested framework and show that extensive, heterogeneous data integration with EL will yield reliable, clinically meaningful mental health predictions.

In summary, this study provides robust empirical evidence that ensemble ML classifiers—particularly RF—achieve clinically meaningful classification performance in multimodal, multisource mental health datasets. The RF classifier’s combination of only 6 FPs among 780 ND samples, a 97.996% overall ACC, and a uniform 0.98 class-weighted F establishes it as the model of choice for deployment in initial-contact DP screening systems. The staged modeling approach confirms a consistent performance hierarchy: DT < linear SVM < boosting ensembles (AdaBoost, XGBoost, LightGBM) < bagging ensemble (RF), affirming the methodological value of data integration across heterogeneous mental health corpora.

These computational results, combined with the epidemiological evidence on the mental health effects of sleep disturbance [59], academic stress [60], and anxiety linked to global crises in young people [58] make the ecological validity of the current framework more profound and underline the clinical urgency of scalable, algorithmically sound early-stage detection systems. With the weight of the burden of youth mental health disorders rising all around the world, AI-assisted classification instruments, when created responsibly, rigorously tested, and applied fairly, provide a major opportunity to narrow the gap between the need and the provision of mental healthcare. Future work will focus on validating this pipeline against clinically annotated longitudinal cohorts, integrating source-held-out

validation strategies to mitigate platform-style leakage, and developing computationally lightweight versions for privacy-preserving on-device deployment.

Future work will focus on strengthening the educational grounds, such as student-related factors, academic stress, and behavioral features, to identify depression risk in students independently of social media, the emotion-corpus dataset, and the workplace. More specifically, SHAP and LIME analyses can quantify the contributions of academic workload, sleeping patterns, and academic pressure to individual predictions; based on these patterns, counselors can interpret why the student is flagged. Longitudinal institutional datasets that link learning-management-system activity, attendance, and grades with stress, anxiety, and depression measures are also needed to test whether risk can be detected before symptoms escalate and to quantify the consequences for academic engagement, performance, and dropout. Finally, deployment in educational settings must attend to student privacy, fairness across gender, programme, and socioeconomic groups, informed consent, and the risk of stigma; the models should be embedded in counselor-supported early-alert workflows in which human judgment remains central, and lightweight, privacy-preserving on-device versions should be explored.

Data Availability Statement

The datasets and source code supporting the findings of this study are publicly available. The datasets can be accessed through the following links: <https://data.mendeley.com/datasets/mmnzx4w8cg/1> (OSMI), <https://www.kaggle.com/datasets/gargmanas/sentimental-analysis-for-tweets?resource=download> (Tweets), <https://www.kaggle.com/datasets/abdullahashfaqvirk/student-mental-health-survey> (Mental Health IT), <https://www.kaggle.com/datasets/hopesb/student-depression-dataset> (Student Depression), <https://data.mendeley.com/datasets/46j8wrc7p7/1> (Health and Sleep), <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health> (Sentiment Analysis), <https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp> (Emotional NLP Dataset), <https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned> (Depression: Reddit), <https://www.kaggle.com/datasets/debarshichanda/goemotions> (GoEmotions), <https://www.kaggle.com/datasets/faisalsanto007/isear-dataset> (ISEAR), and <https://www.kaggle.com/datasets/dartweichen/student-life> (Student Life). The implementation details and source code are available at: <https://github.com/RamSitha/AI-YMH>.

Funding

This work was supported by Telecommunications Consultants India Limited (TCIL), a Government of India enterprise under the administrative control of the Department of Telecommunications (DoT), as part of the “100 5G Use Case Labs” initiative.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

- [1] Uhlhaas, P. J., Davey, C. G., Mehta, U. M., Shah, J., Torous, J., Allen, N. B., ... & Wood, S. J. (2023). Towards a youth mental health paradigm: a perspective and roadmap. *Molecular psychiatry*, 28(8), 3171-3181. [CrossRef]
- [2] Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Üstün, T. B. (2007). Age of onset of mental disorders: a review of recent literature. *Current Opinion in Psychiatry*, 20(4), 359-364. [CrossRef]
- [3] McGorry, P. D., & Mei, C. (2018). Early intervention in youth mental health: progress and future directions. *Evidence-Based Mental Health*, 21(4), 182-184. [CrossRef]
- [4] Treatment for Adolescents with Depression Study (TADS) Team. (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents with Depression Study (TADS) randomized controlled trial. *Jama*, 292(7), 807-820. [CrossRef]
- [5] Vyas, N. S., Birchwood, M., & Singh, S. P. (2015). Youth services: meeting the mental health needs of adolescents. *Irish Journal of Psychological Medicine*, 32(1), 13-19. [CrossRef]
- [6] McGorry, P., Trethowan, J., & Rickwood, D. (2019). Creating headspace for integrated youth mental health care. *World Psychiatry*, 18(2), 140-141. [CrossRef]
- [7] Yung, A. R., Yung, A. R., Pan Yuen, H., McGorry, P. D., Phillips, L. J., Kelly, D., ... & Buckby, J. (2005). Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. *Australian & New Zealand Journal of Psychiatry*, 39(11-12), 964-971. [CrossRef]
- [8] Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: a meta-analysis. *Psychological Bulletin*, 132(1), 132-149. [CrossRef]
- [9] Appleton, R., Barnett, P., Clarke, C., Yang, J., Begum, S., Edbrooke-Childs, J., ... & Johnson, S. (2025). Approaches to early intervention for common mental health problems in young people: a systematic review. *BMC medicine*, 23(1), 651. [CrossRef]
- [10] Torous, J., Myrick, K. J., Rauseo-Ricupero, N., & Firth, J. (2020). Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR mental health*, 7(3), e18848. [CrossRef]
- [11] Domhardt, M., Steubl, L., & Baumeister, H. (2018). Internet-and mobile-based interventions for mental and somatic conditions in children and adolescents. *Zeitschrift für Kinder-und Jugendpsychiatrie und Psychotherapie*, 48(1), 33-46. [CrossRef]
- [12] Ho, T. Q. A., Le, L. K.-D., Engel, L., Le, N., Melvin, G., Le, H. N. D., & Mihalopoulos, C. (2025). Barriers to and facilitators of user engagement with web-based mental health interventions in young people: a systematic review. *European Child & Adolescent Psychiatry*, 34(1), 83-100. [CrossRef]
- [13] Nesi, J. (2020). The impact of social media on youth mental health: challenges and opportunities. *North Carolina Medical Journal*, 81(2), 116-121. [CrossRef]
- [14] Power, E., Hughes, S., Cotter, D., & Cannon, M. (2020). Youth mental health in the time of COVID-19. *Irish Journal of Psychological Medicine*, 37(4), 301-305. [CrossRef]
- [15] Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, 47(1), 1-18. [CrossRef]
- [16] McGorry, P. D., Mei, C., Chanen, A., Hodges, C., Alvarez-Jimenez, M., & Killackey, E. (2022). Designing and scaling up integrated youth mental health care. *World Psychiatry*, 21(1), 61-76. [CrossRef]
- [17] Marshall, N. J., Loades, M. E., Jacobs, C., Biddle, L., & Lambert, J. D. (2025). Integrating artificial intelligence in youth mental health care: Advances, challenges, and future directions. *Current Treatment Options in Psychiatry*, 12(1), 11. [CrossRef]
- [18] Mlakar, I., Šafran, V., Plohl, N., Kumperščak, H. G., Močnik, S., Smogavc, Ž., & Smrke, U. (2026). Technology-mediated screening interviews for youth mental health: content validation, randomized controlled trial, and expert evaluation. *PLOS Digital Health*, 5(4), e0001069. [CrossRef]
- [19] Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: a survey of

- current progress in depression detection, diagnosis and treatment. *Brain Informatics*, 10(1), 10. [CrossRef]
- [20] Agarwal, N., Dias, G., & Dollfus, S. (2024). Multi-view graph-based interview representation to improve depression level estimation. *Brain Informatics*, 11(1), 14. [CrossRef]
- [21] Sharma, G., Yaffe, M. J., Ghadiri, P., Gandhi, R., Pinkham, L., Gore, G., & Abbasgholizadeh-Rahimi, S. (2025). Use of artificial intelligence in adolescents' mental health care: systematic scoping review of current applications and future directions. *JMIR Mental Health*, 12, e70438. [CrossRef]
- [22] Zhou, Q., Chen, Z. H., Cao, Y. H., & Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ digital medicine*, 4(1), 154. [CrossRef]
- [23] Chang, J. P. C., Cheng, S. W., Chang, S. M. J., & Su, K. P. (2025). Navigating the digital maze: a review of AI bias, social media, and mental health in Generation Z. *AI*, 6(6), 118. [CrossRef]
- [24] Koutsouleris, N., Hauser, T. U., Skvortsova, V., & De Choudhury, M. (2022). From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health*, 4(11), e829-e840. [CrossRef]
- [25] Arias, D., Saxena, S., & Verguet, S. (2022). Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine*, 54, 101675. [CrossRef]
- [26] Maughan, B., Collishaw, S., & Stringaris, A. (2013). Depression in childhood and adolescence. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 22(1), 35-40. <https://kclpure.kcl.ac.uk/portal/en/publications/depression-in-childhood-and-adolescence>
- [27] Torous, J., & Blease, C. (2024). Generative artificial intelligence in mental health care: potential benefits and current challenges. *World Psychiatry*, 23(1), 1-2. [CrossRef]
- [28] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2), e19. [CrossRef]
- [29] Gopalakrishnan, A., Gururajan, R., Venkataraman, R., Zhou, X., Ching, K. C., Saravanan, A., & Sen, M. (2023). Attribute selection hybrid network model for risk factors analysis of postpartum depression using social media. *Brain Informatics*, 10(1), 28. [CrossRef]
- [30] Moor, S., Williman, J., Drummond, S., Fulton, C., Mayes, W., Ward, N., ... & Stasiak, K. (2019). 'E'therapy in the community: Examination of the uptake and effectiveness of BRAVE (a self-help computer programme for anxiety in children and adolescents) in primary care. *Internet Interventions*, 18, 100249. [CrossRef]
- [31] Gohari, M. R., Doggett, A., Patte, K. A., Ferro, M. A., Dubin, J. A., Hilario, C., & Leatherdale, S. T. (2024). Using random forest to identify correlates of depression symptoms among adolescents. *Social psychiatry and psychiatric epidemiology*, 59(11), 2063-2071. [CrossRef]
- [32] Mullick, T., Radovic, A., Shaaban, S., & Doryab, A. (2022). Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. *JMIR Formative Research*, 6(6), e35807. [CrossRef]
- [33] Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., ... & Baumeister, H. (2018). Internet-and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. *European Psychologist*, 23(2), 167-187. [CrossRef]
- [34] Grist, R., Croker, A., Denne, M., & Stallard, P. (2019). Technology delivered interventions for depression and anxiety in children and adolescents: a systematic review and meta-analysis. *Clinical child and family psychology review*, 22(2), 147-171. [CrossRef]
- [35] Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., & Nguyen, Q. V. (2019). What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. *Frontiers in Psychiatry*, 10, 759. [CrossRef]
- [36] Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental Review*, 28(1), 78-106. [CrossRef]
- [37] Barker, E. T., Howard, A. L., Villemare-Krajden, R., & Galambos, N. L. (2018). The rise and fall of depressive symptoms and academic stress in two samples of university students. *Journal of Youth and Adolescence*, 47(6), 1252-1266. [CrossRef]
- [38] Bhujade, V. M. (2017). Depression, anxiety and academic stress among college students: A brief review. *Indian Journal of Health and Wellbeing*, 8(7), 748-751. <https://www.proquest.com/openview/62b3afd412639d502f219859b188a281/1?pq-origsite=gscholar&cbl=2032134>
- [39] Veal, C., Tomlinson, A., Cipriani, A., Bulteau, S., Henry, C., Müh, C., ... & Chevance, A. (2024). Heterogeneity of outcome measures in depression trials and the relevance of the content of outcome measures to patients: a systematic review. *The Lancet Psychiatry*, 11(4), 285-294. [CrossRef]
- [40] Andargeery, S. Y. (2024). The perception and association between depression and academic stress among female undergraduate nursing students: a cross-sectional study. *Frontiers in Public Health*, 12, 1414469. [CrossRef]
- [41] Peverill, M., Dirks, M. A., Narvaja, T., Herts, K. L., Comer, J. S., & McLaughlin, K. A. (2021).

- Socioeconomic status and child psychopathology in the United States: A meta-analysis of population-based studies. *Clinical psychology review*, 83, 101933. [CrossRef]
- [42] Kemp, J., Chorney, J., Kassam, I., MacDonald, J., MacDonald, T., Wozney, L., & Strudwick, G. (2021). Learning about the current state of digital mental health interventions for Canadian youth to inform future decision-making: mixed methods study. *Journal of Medical Internet Research*, 23(10), e30491. [CrossRef]
- [43] Zhai, Y., Zhang, Y., Chu, Z., Geng, B., Almaawali, M., Fulmer, R., ... & Du, X. (2025). Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students. *Journal of Counseling & Development*, 103(1), 110-125. [CrossRef]
- [44] Graham, A. K., Lattie, E. G., Powell, B. J., Lyon, A. R., Smith, J. D., Schueller, S. M., ... & Mohr, D. C. (2020). Implementation strategies for digital mental health interventions in health care settings. *American Psychologist*, 75(8), 1080. [CrossRef]
- [45] Langley, A. K., Nadeem, E., Kataoka, S. H., Stein, B. D., & Jaycox, L. H. (2010). Evidence-based mental health programs in schools: Barriers and facilitators of successful implementation. *School Mental Health*, 2(3), 105-113. [CrossRef]
- [46] Rojas-Andrade, R., & Bahamondes, L. L. (2019). Is implementation fidelity important? A systematic review on school-based mental health programs. *Contemporary School Psychology*, 23(4), 339-350. [CrossRef]
- [47] Zhang, Y., Larson, M., Ehrhart, M. G., King, K., Locke, J., Cook, C. R., & Lyon, A. R. (2024). Inter-organizational alignment and implementation outcomes in integrated mental healthcare for children and adolescents: a cross-sectional observational study. *Implementation Science*, 19(1), 36. [CrossRef]
- [48] Shalaby, R. A. H., & Agyapong, V. I. (2020). Peer support in mental health: literature review. *JMIR mental health*, 7(6), e15572. [CrossRef]
- [49] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195. [CrossRef]
- [50] Powell, A. C., Bowman, M. B., & Harbin, H. T. (2019). Reimbursement of apps for mental health: findings from interviews. *JMIR mental health*, 6(8), e14724. [CrossRef]
- [51] Zhao, X., Stadnick, N. A., Ceballos-Corro, E., Castro Jr, J., Mallard-Swanson, K., Palomares, K. J., ... & Sorkin, D. H. (2023). Facilitators of and barriers to integrating digital mental health into county mental health services: qualitative interview analyses. *JMIR Formative Research*, 7, e45718. [CrossRef]
- [52] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106. [CrossRef]
- [53] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297. [CrossRef]
- [54] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. [CrossRef]
- [55] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139. [CrossRef]
- [56] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.
- [57] 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). [CrossRef]
- [58] Kaman, P. D. A., Devine, P. D. J., Erhart, P. D. M., Napp, M. S. A. K., Reiss, P. D. F., Moeller, S., ... & Ravens-Sieberer, P. D. U. (2025). Youth mental health in times of global crises: evidence from the German longitudinal COVID-19 and psychological health study. *Journal of Adolescent Health*, 77(6), 1050-1057. [CrossRef]
- [59] Vestergaard, C. L., Skogen, J. C., Hysing, M., Harvey, A. G., Vedaa, Ø., & Sivertsen, B. (2024). Sleep duration and mental health in young adults. *Sleep Medicine*, 115, 30-38. [CrossRef]
- [60] Kar, B., & Sarkar, B. K. (2024). Changes in mental health of Indian students due to online classes during COVID-19 pandemic. *Procedia Computer Science*, 235, 2452-2465. [CrossRef]
- [61] Rasheed, F., Khan, S., Waqas, S. M., Khan, R. A., & Arif, S. (2024). Survey dataset on mental health in tech professionals from open sourcing mental health surveys (2017-2021). *Data in Brief*, 54, 110377. [CrossRef]



Ramesh K Bhukya received the B.Tech. degree in Electronics and Communication Engineering from Mahatma Gandhi Institute of Technology (MGIT), Gandipet, Hyderabad, Telangana, India, in 2008, the M.Tech. degree in Electronics and Communication Engineering from Motilal Nehru National Institute of Technology (MNNIT), Allahabad, Prayagraj, Uttar Pradesh, India, in 2011, and the Ph.D. degree in Electronics and Electrical

Engineering from Indian Institute of Technology Guwahati (IITG), Guwahati, Assam, India, in 2019. He joined as an Assistant Professor in Electronics and Communication Engineering at the National Institute of Technology (NIT), Hamirpur, Himachal Pradesh, India, in 2019-2020. In 2020, he joined the Department of Electronics and Communication Engineering at National Institute of Technology Andhra Pradesh (NIT-AP), Tadepalligudem, Andhra Pradesh, India. He is currently working as an Assistant Professor in the Department of Electronics and Communication Engineering at the Indian Institute of Information Technology Allahabad (IIIT-A), Prayagraj, Uttar Pradesh, India. His research

interests are speech signal processing, speaker verification, biomedical signal processing, machine learning, and pattern recognition. (Email: rkbhukya@iiita.ac.in)



Nirbhay Singh has completed his undergraduate studies with a Bachelor of Technology in the Department of Electronics and Communication Engineering at Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India. He has a data science degree certificate issued by IIT Madras. His research areas are biomedical signal processing, speech technology, machine learning, deep learning, and computer vision, as well as their applications in intelligent diagnostic systems. He is also undergoing an AI/ML Internship in the Department of Electronics and Communication Engineering at the Indian

Institute of Information Technology, Allahabad, Prayagraj-211015, Uttar Pradesh, India. (Email: nirbhaysinghhro@gmail.com)



Aniket Tiwari has received a Bachelor of Technology in the Department of Computer Science and Engineering at the United College of Engineering and Research, UPSIDC Industrial Area, Naini, Prayagraj, Uttar Pradesh. His research interests include speech processing, machine learning, deep learning, and computer vision, with a particular focus on their applications in intelligent drone surveillance systems. He has undertaken an AI-ML Internship in the Department of Electronics and Communication Engineering at the Indian Institute of Information Technology Allahabad (IIIT-A), Prayagraj, Uttar Pradesh, India. (Email: aniketiwari007@gmail.com)