Temporal Validation of Machine Learning Models Utilized for Mental Health Prediction in College Students: A Three-Year Longitudinal Study

Tolkyn Tuleutayeva
School of Software Engineering, Astana IT University, Astana, Kazakhstan
ttuleutayeva@outlook.com

Alua Myrzakerimova
School of Software Engineering, Astana IT University, Astana, Kazakhstan
Alua.Myrzakerimova@astanait.edu.kz (corresponding author)

M. O. Nurmaganbetova
Department of Medical Biophysics, Informatics, and Mathematical Statistics, Kazakh National Medical University named after S. Asfendiyarov, Almaty, Kazakhstan
mug2009@mail.ru

N. Nalgozhina
Software Engineering Department, Satbayev University, Almaty, Kazakhstan
n.nalgozhina@satbayev.university

Received: 22 January 2026 | Revised: 21 February 2026 and 25 March 2026 | Accepted: 27 March 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.17715*

ABSTRACT

This study investigates the longitudinal robustness of mental health prediction models using three years of data from the Healthy Minds Study, comprising 265,870 college students surveyed between 2022 and 2025. Multiple statistical techniques were applied to assess data drift, while three Machine Learning (ML) algorithms, namely Logistic Regression (LR), Random Forest (RF), and XGBoost, were evaluated under several temporal modeling strategies. Mental health risk was defined as the presence of moderate-to-severe depression or anxiety symptoms. Although statistically significant distributional changes were observed across variables, effect sizes remained small, indicating limited practical drift. Model performance remained strong over time (mean F1-score = 0.71, mean AUROC = 0.80), with minimal temporal degradation (1.8%). Well-being emerged as the most influential predictor, accounting for the dominant share of feature importance. These findings suggest that mental health prediction models can be reliably deployed when temporal stability is present and highlight the crucial role of well-being in mental health risk prediction.

Keywords-machine learning; mental health; temporal validation; data drift; college students; depression; anxiety; well-being

I. INTRODUCTION

National longitudinal data indicate that the mental health burden among college students is clinically substantial. Specifically, the 2024-2025 Healthy Minds Study (HMS), conducted across 135 U.S. institutions with over 84,000 respondents, revealed that 37% of students reported moderate-to-severe depressive symptoms, whereas 32% reported moderate-to-severe anxiety [1]. Although these figures represent a third consecutive year of decline from pandemic-era peaks (44% and 37%, respectively), they remain significantly elevated compared to pre-pandemic baselines [2].

Traditional mental health care in college settings relies primarily on self-referral. However, this reactive model fails to reach most students who need help, with only 20-40% of those experiencing mental health problems actually seeking treatment [3, 4]. This gap between need and care has motivated

the interest in proactive screening approaches, with Machine Learning (ML) being a potentially scalable solution [5, 6].

The transition from promising research results to reliable real-world deployment faces a significant challenge that has received little attention: temporal data drift [7, 8]. Data drift refers to systematic changes in data distributions over time; that is, shifts that can silently degrade model performance if left unaddressed. In healthcare, drift may be induced by shifting demographics, changing measurement protocols, evolving awareness of symptoms, or major societal events [9, 10]. The COVID-19 pandemic, for example, altered mental health patterns across populations [11, 12]. The post-pandemic recovery period has introduced its own distinct dynamic: as acute crisis conditions have receded, students are navigating a gradual rebuilding of social connectedness and well-being, yet resilience trajectories vary substantially across demographic groups and institutional contexts.

These shifts in well-being and resilience during recovery represent a meaningful source of distributional change that ML models trained on pre- or early-pandemic data may not adequately capture. Beyond large-scale disruptions, even subtle year-to-year shifts in student demographics, institutional composition, or societal stressors can progressively erode model reliability, often without triggering obvious performance alerts. This makes prospective drift monitoring not merely a methodological refinement, but an operational necessity for any deployed screening system.

Despite the potential importance of temporal drift, most mental health ML studies evaluate models using cross-sectional data or report performance on held-out test sets from the same time period as training data [13, 14]. While these approaches establish that models can learn meaningful patterns, they provide limited insights into whether models maintain performance over time, a significant requirement for operational deployment.

The present study addresses this gap by comprehensively validating ML models using three consecutive years of HMS data (2022-2025). Three key objectives were investigated: (1) to what extent does practical data drift occur in longitudinal college mental health data, as measured through effect-size-based metrics, (2) how does temporal drift impact the predictive performance of ML models when deployed across different time periods, and (3) what modeling strategies maintain stability in the presence of temporal changes.

Unlike prior ML studies in college mental health that relied primarily on cross-sectional validation or random train-test splits within a single cohort [13, 14], the present study performs explicit multi-year temporal validation across independent annual samples. By integrating formal drift detection techniques, including effect size interpretation, Population Stability Index (PSI) assessment, and correlation stability analysis, with cross-year predictive evaluation, this research directly examines real-world deployment robustness rather than isolated predictive accuracy. This distinction is essential for translating ML research into sustainable operational systems.

II. METHODS

A. Data Source and Participants

Data from the HMS, an annual web-based survey that assesses mental health, service utilization, and related factors among college students at participating U.S. institutions, were analyzed in this study [1]. HMS employs stratified random sampling within institutions, drawing a random sample of enrolled students from institutional registries and inviting them via email. Participation is voluntary and anonymous, with typical response rates between 15-30%, consistent with norms for institutional web-based surveys. The HMS sampling frame spans both two-year and four-year institutions, though four-year research universities constitute the majority of participating institutions. Inclusion criteria for the present analysis were: (1) currently enrolled undergraduate or graduate students at a participating U.S. institution; (2) valid responses on at least the PHQ-9, GAD-7, and Flourishing Scale; and (3) survey completion in one of the three target academic years (2022-2023, 2023-2024, or 2024-2025). Students with entirely missing outcome data were excluded prior to analysis. The combined analytic sample included 265,870 students across three consecutive academic years: 76,406 students in 2022-2023 (28.7%), 104,729 students in 2023-2024 (39.4%), and 84,735 students in 2024-2025 (31.9%). These data span approximately 200 institutions across all major U.S. geographic regions (Northeast, Southeast, Midwest, Southwest, and West), including public and private institutions of varying size and selectivity, providing a nationally representative sample of U.S. college students.

B. Mental Health Measures

Mental health risk was operationalized as the presence of moderate-to-severe symptoms of depression or anxiety - symptom levels that typically warrant clinical intervention [15, 16]. Depression was assessed using the Patient Health Questionnaire-9 (PHQ-9), a well-validated 9-item instrument that measures depressive symptoms over the past two weeks [15]. Scores ranged from 0 to 27, with values ≥ 10 indicating moderate-to-severe depression. On the other hand, anxiety was measured using the Generalized Anxiety Disorder-7 (GAD-7), which is a 7-item scale assessing anxiety symptoms over the past two weeks [16]. Scores varied from 0 to 21, with values ≥ 10 showing moderate-to-severe anxiety. A student was classified as high risk (Positive class, label = 1) if PHQ-9 ≥ 10 or GAD-7 ≥ 10. In the analytic sample of 234,544 students, 111,449 (47.5%) were Positive and 123,095 (52.5%) were Negative, reflecting a near-balanced class distribution that required no oversampling or class-weight correction.

Well-being was assessed using the Flourishing Scale, an 8-item measure of psychological well-being covering domains including purpose, relationships, engagement, and self-acceptance [17]. Scores ranged from 8 to 56, with higher values indicating greater well-being. Predictor selection was constrained by missingness patterns in the data. To ensure that the models could be applied reliably in practice, the predictors were limited to those with less than 30% missing data. This resulted in six predictors: Flourishing Scale total score (continuous, 1.35% missing), loneliness (binary: yes/no, 1.43% missing), age (continuous, no missing data), food insecurity

worry (3-point ordinal scale, 0.95% missing), housing insecurity worry (3-point ordinal scale, 0.12% missing), and international student status (binary, no missing data).

The 30% missingness threshold was selected to balance predictor completeness with retention of an adequate sample size for stable model estimation. Given the large overall sample (N > 250,000), prioritizing predictors with higher completeness minimized potential bias introduced by extensive imputation procedures and reflected realistic deployment conditions in which complex imputation pipelines may not be operationally feasible.

TABLE I. PREDICTOR VARIABLES AND MISSINGNESS

Predictor	Type	Missing (%)
Flourishing Scale (total score)	Continuous	1.35
Loneliness	Binary	1.43
Age	Continuous	0.0
Food insecurity worry	Ordinal (1-3)	0.95
Housing insecurity worry	Ordinal (1-3)	0.12
International student status	Binary	0.0

C. Drift Detection Methods

In order to detect temporal drift, multiple complementary approaches were employed, with a focus on the effect sizes rather than p-values, as due to the large sample, there is a statistical power to detect extraordinarily small differences (differences that may be statistically significant but practically meaningless) [18]. For continuous variables, two-sample Kolmogorov-Smirnov (K-S) tests were applied to detect distributional differences between year pairs [19]. These tests were complemented with Cohen's d effect sizes for mean differences, using established interpretation guidelines: d < 0.20 as trivial effects, 0.20-0.50 as medium effects, and > 0.50 as large effects [20].

PSI quantifies distributional shift by comparing the distribution of a variable across discrete bins [21]. Its 0.10 threshold for stability is widely adopted in credit risk modeling and ML deployment monitoring [21]. Specifically, values < 0.10 indicate distributions stable enough that models require no immediate recalibration; values of 0.10-0.25 suggest that monitoring is needed, and values > 0.25 indicate significant drift warranting model retraining or recalibration. For categorical variables, chi-square tests combined with Cramér's V effect sizes were utilized. Except for these, correlation matrices were examined across years to assess whether relationships among predictors remained stable, flagging changes of 0.10 or more in absolute correlation as potentially meaningful, consistent with guidelines for practical significance in psychological measurement research.

D. Machine Learning Models

Three ML algorithms were evaluated, representing different modeling approaches. Logistic Regression (LR) with L2 regularization provided an interpretable baseline, Random Forest (RF) with 100 trees and a maximum depth of 10 captured nonlinear relationships, while XGBoost with 100 boosting rounds, a maximum depth of 6, and a learning rate of 0.1 represented a state-of-the-art gradient boosting approach [22, 23]. Missing values were imputed using median imputation for continuous variables and mode imputation for categorical ones. Median/mode imputation was selected as a conservative and reproducible approach suitable for large-scale deployment settings. Although more complex imputation strategies may be appropriate under certain assumptions about missingness, they introduce additional modeling complexity that may not be necessary in high-sample contexts. Continuous features were standardized before model training. To prevent data leakage, all preprocessing steps, including imputation and standardization, were encapsulated within scikit-learn pipelines, fitted exclusively on training data, and applied independently to each test set. Hyperparameters were selected using standard default configurations without extensive grid search optimization, as the primary objective of this study was to assess temporal robustness rather than maximize peak predictive performance. Table II summarizes the hyperparameter configurations used in all models.

TABLE II. HYPERPARAMETER CONFIGURATIONS

Model	Key hyperparameters	Tuning strategy
LR	L2 regularization, C = 1.0	Default configuration
RF	n_estimators = 100, max_depth = 10, min_samples_split = 2, min_samples_leaf = 1	Default configuration
XGBoost	n_estimators = 100, max_depth = 6, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 0.8	Default configuration

Four temporal modeling strategies were compared in order to identify approaches that maintain performance across time. Strategy 1 (Baseline) trained separate models for each year using 80/20 train-test splits within that year. Strategy 2 (Cross-Year Validation) trained models on one year and tested on all three years, directly quantifying temporal performance degradation. Strategy 3 (Pooled Model) trained a single model on combined data from all three years. Strategy 4 (Temporal Feature Model) extended the pooled model by including academic year as an explicit predictor.

Model performance was evaluated using three metrics. F1 score served as the primary metric because it balances concerns about false positives and false negatives. AUC-ROC assessed discrimination ability across classification thresholds. Accuracy provided an overall performance measure. All analyses were conducted using Python 3.10 with scikit-learn, XGBoost, SciPy, and visualization libraries.

III. RESULTS

A. Sample Characteristics and Trends

The initial sample comprised 265,870 students across three academic years. After excluding 31,326 respondents (11.8%) with missing outcome data, the final sample contained 234,544 students (67,961 in 2022-2023, 94,086 in 2023-2024, and 72,497 in 2024-2025). The sample was predominantly female (68%) with a mean age of 23.7 years (SD = 7.6), distributed across 176 participating institutions. International students made up approximately 7-8% of the sample, and approximately 50-55% of students reported experiencing loneliness, with rates declining across the three years.

TABLE III. MENTAL HEALTH TRENDS ACROSS THREE ACADEMIC YEARS

Measure	2022-2023	2023-2024	2024-2025	Cohen's d
Depression (PHQ-9)	9.05 (6.40)	8.89 (6.30)	8.36 (6.19)	-0.111
Moderate + Depression (%)	40.7	40.0	36.8	-
Anxiety (GAD-7)	8.14 (5.87)	7.95 (5.87)	7.53 (5.77)	-0.105
Moderate + Anxiety (%)	37.1	35.6	33.4	-
Well-Being	43.06 (9.00)	43.22 (9.00)	43.53 (9.03)	+0.053

Mental health trends during the post-pandemic period were generally favorable, though the overall prevalence of clinically significant symptoms remained high. Specifically, depression scores (PHQ-9) dropped from 9.05 in 2022-2023 to 8.36 in 2024-2025, with moderate-to-severe depression prevalence declining from 40.7% to 36.8%. Similarly, anxiety scores (GAD-7) decreased from 8.14 to 7.53, with moderate-to-severe anxiety prevalence decreasing from 37.1% to 33.4% over the same period. Well-being scores presented modest improvement, increasing from 43.06 to 43.53. Table III summarizes these trends across all three academic years.

B. Drift Detection Results

Drift testing showed a clear pattern with statistical tests detecting substantial differences in distributions. However, the effect sizes remained at minimal levels. Every K-S test, which compared data from 2022-2023 with data from 2024-2025, produced *p*-values under 0.001, while Cohen's d effect sizes only ranged from 0.001 to 0.115, which are below the 0.20 level that researchers consider to be significant. PSI values demonstrated distributional stability as they remained below the 0.10 threshold. Correlation matrices across academic years further confirmed the stability of the relationships among predictors (Figure 1).

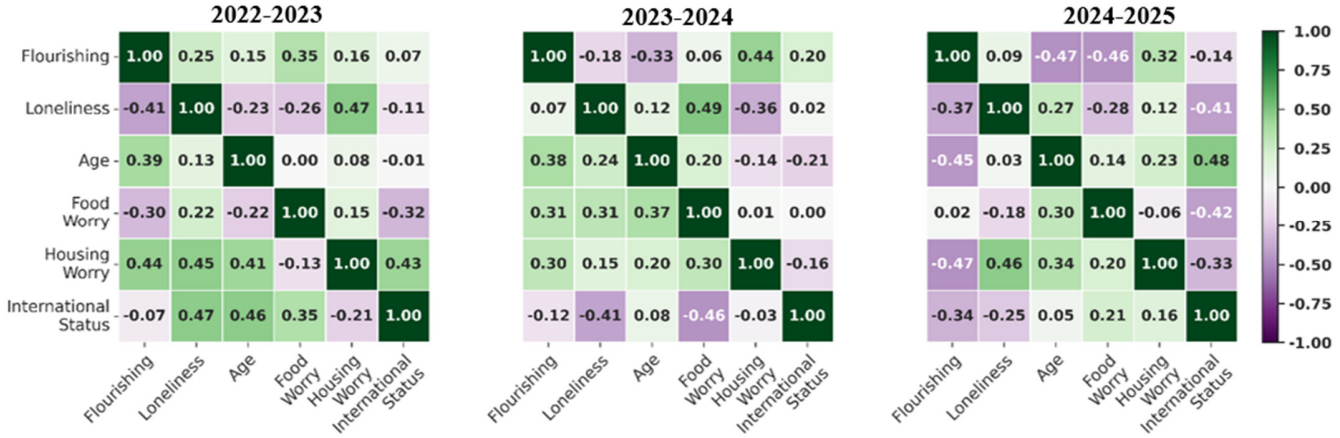

Fig. 1. Correlation matrices across academic years (2022–2025) showing stability of relationships among predictors. Values represent Pearson correlation coefficients.

The maximum absolute change in any pairwise correlation was 0.041, which is below the pre-specified threshold of 0.10. These patterns indicated that not only did individual variable distributions remain stable, but the underlying relationships among predictors also stayed consistent. For instance, the negative correlation between flourishing and loneliness did not change ($-0.39 \leq r \leq -0.41$), while the correlation between food and housing insecurity worry varied only slightly ($0.43 \leq r \leq 0.50$). Table IV summarizes drift detection results.

TABLE IV. DRIFT DETECTION SUMMARY

Variable	K-S Test	Effect Size	PSI
PHQ-9	$p < 0.001$	$d = -0.111$	0.013
GAD-7	$p < 0.001$	$d = -0.105$	0.012
Flourishing	$p < 0.001$	$d = +0.053$	0.005
Age	$p < 0.001$	$d = +0.115$	0.015
Food worry	$p = 1.000$	$d = -0.001$	0.000
Housing worry	$p = 0.475$	$d = +0.012$	0.000
Loneliness	$p < 0.001$	$V = 0.040$	0.000

C. Machine Learning Model Performance

ML models achieved strong performance that remained stable across all three years. RF emerged as the top performer, achieving a mean *F1* score of 0.708 across the three years, with *AUC* values averaging 0.801. XGBoost performed nearly identically, also achieving *F1* = 0.711. The year-to-year variability in *F1* scores was remarkably small, with a range of only 0.033. LR reached respectable but noticeably lower performance (*F1* = 0.680 – 0.7130, *AUC* =0.784 - 0.796 across baseline per-year evaluations), suggesting that the relationships between predictors and mental health risk involve nonlinear patterns that linear models cannot fully capture. The stability of model performance across years is particularly noteworthy given the substantial sample sizes involved. With over 76,000 students in the smallest annual cohort and over 104,000 in the largest, random fluctuations in model performance would be minimal, making the observed consistency even more meaningful. The negligible difference between RF and XGBoost performance (RF: *F1* = 0.708, XGB: *F1* = 0.711) suggests that the underlying patterns in data are robust enough

that different modeling approaches converge to similar solutions. This finding has practical implications: organizations can select models based on operational considerations, such as interpretability and computational efficiency, rather than pursuing marginal performance gains through algorithmic complexity. The superior performance of tree-based ensemble methods (RF, XGBoost) compared to LR points to important nonlinearities in the relationships between predictors and mental health risk. These nonlinearities likely reflect interaction effects. For example, the protective effect of well-being may vary depending on the presence or absence of loneliness, or the impact of housing insecurity may differ across age groups. Cross-year validation provided the most direct test of temporal robustness. When RF models were trained on one year and tested on different years, the performance degradation was minimal. Models tested on their training year achieved an *F1* score equal to 0.724, while when tested on different years, they reached *F1* = 0.711 - a degradation of only 1.8%. This strong temporal stability indicates that patterns learned from one year's data generalize well to subsequent years (Figure 2).

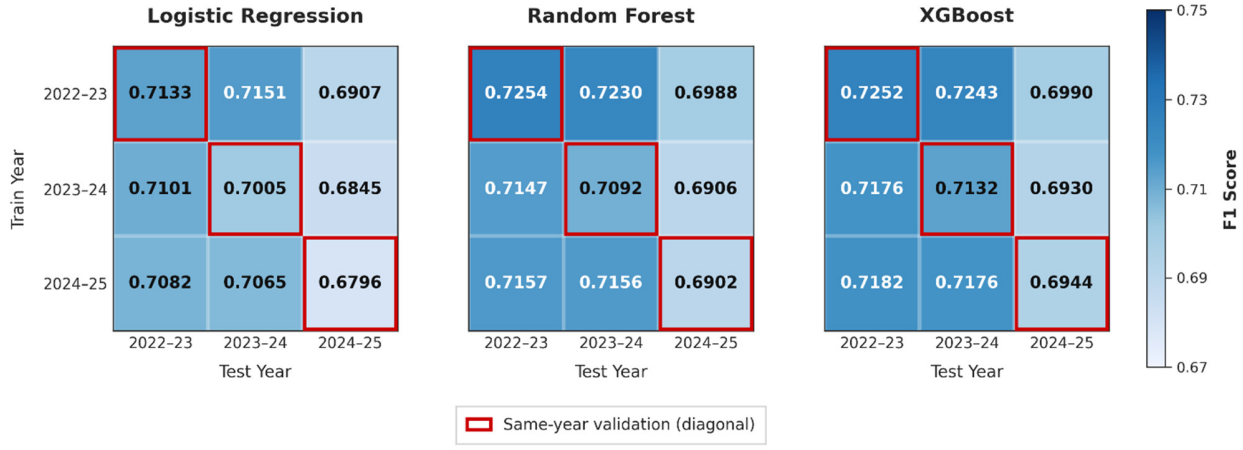

Fig. 2. Cross-year validation results showing F1 scores by training and test year for LR, RF, and XGBoost models. Diagonal cells (red outline) represent same-year validation; off-diagonal cells reflect cross-year generalization performance.

The baseline model performance across all three academic years is displayed in Figure 3. RF and XGBoost demonstrated consistently high performance across all metrics, with F1 scores of 0.69-0.73 and AUC-ROC values of 0.79-0.81 in every year. LR showed noticeably lower performance, particularly for AUC-ROC (0.784-0.796), highlighting the importance of nonlinear modeling approaches for capturing complex relationships between predictors and mental health outcomes. The stability of performance metrics across years further confirmed the temporal robustness observed in cross-year validation analyses. This consistency was maintained despite the declining prevalence of high-risk cases from 49.5% in 2022-2023 to 45.2% in 2024-2025, suggesting that models trained on earlier cohorts remain well-calibrated for more recent populations.

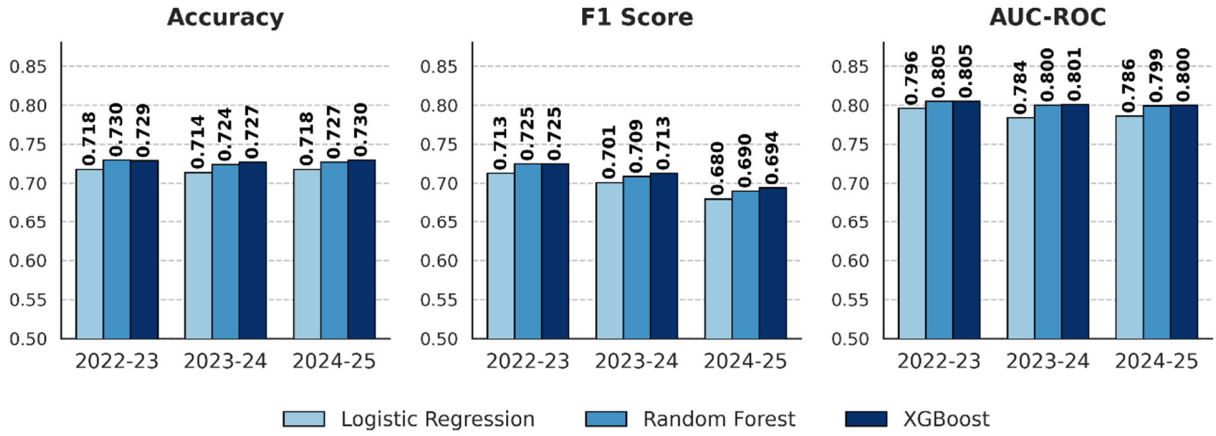

Fig. 3. Baseline model performance across academic years comparing LR, RF, and XGBoost using Accuracy, F1 Score, and AUC-ROC.

Table V presents the detailed performance results across all temporal validation strategies. Confusion matrices for all three models evaluated on the pooled test set are reported in Table VI.

TABLE V. MODEL PERFORMANCE ACROSS TEMPORAL STRATEGIES

Strategy	Model	F1 Score	AUC-ROC
Baseline (per-year)	RF	0.708	0.801
Baseline (per-year)	XGBoost	0.711	0.802
Cross-year (same)	RF	0.724	—
Cross-year (different)	RF	0.711	—
Pooled model	RF	0.717	0.807
Pooled + year	RF	0.716	0.807

TABLE VI. CONFUSION MATRICES - POOLED TEST SET (N = 46,909; POSITIVE = 22,290; NEGATIVE = 24,619)

Model	TP	FP	FN	TN	Sensitivity	Specificity
LR	15,653	6,286	6,637	18,333	0.702	0.745
RF	15,826	6,058	6,464	18,561	0.710	0.754
XGBoost	15,803	5,959	6,487	18,660	0.709	0.758

D. Feature Importance Analysis

Feature importance analysis revealed a striking dominance of well-being in predicting mental health risk. The Flourishing Scale accounted for 54.30% of the total feature importance (Figure 4) in the RF model - approximately 1.9 times more important than loneliness. Loneliness ranked second at 28.56%, followed by food insecurity worry (8.70%), age (4.18%), housing insecurity worry (3.74%), and international student status (0.52%).

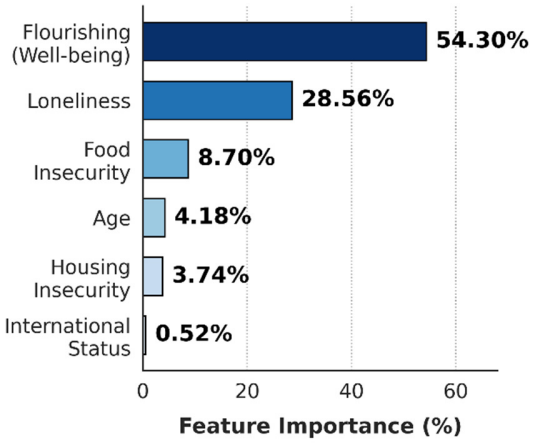

Fig. 4. Feature importance for mental health risk prediction based on mean decrease in Gini impurity (RF, pooled sample). The Flourishing score demonstrates a dominant predictive contribution.

IV. DISCUSSION

The present findings advanced the understanding of ML deployment in mental health contexts along two complementary dimensions: the characterization of temporal data drift and the evaluation of predictive robustness across independent annual cohorts. Considering both of them indicates that well-constructed ML models, trained on longitudinal college mental health data, can maintain strong performance without frequent retraining, provided that distributional monitoring is in place.

A. Statistical Versus Practical Significance

Perhaps the most important methodological finding is the stark disconnect between statistical and practical importance in large samples. Every drift detection test achieved *p*-values below 0.001, yet effect sizes told a completely different story: all fell below 0.15, and PSI values remained well below the 0.10 stability threshold. With 265,870 participants, this research possessed sufficient statistical power to detect truly minuscule effects - so small differences that would have no practical consequence for model performance. This enormous power renders *p*-values nearly useless for assessing practical importance [18]. Consequently, in large sample studies, effect size reporting is essential for distinguishing signal from noise [24].

B. Implications for Model Deployment

The minimal practical drift that was observed has direct implications of how mental health prediction models should be deployed. Organizations can use simple static models trained on historical data without implementing complex continuous retraining systems. Pooled training across multiple time periods maximizes sample size and statistical stability while maintaining operational simplicity. Models trained on 2022-2023 data performed nearly identically when applied to 2024-2025 data, suggesting that annual retraining may be unnecessary if monitoring confirms ongoing stability [25]. However, continuous monitoring remains essential even when historical data show stability [26]. The minimal temporal degradation observed (1.80%) indicated that, under conditions of limited practical drift, periodic monitoring of distributional metrics may be sufficient to maintain model reliability. In such contexts, continuous large-scale retraining might not be necessary, reducing operational complexity while preserving predictive stability.

C. Generalizability and Response Bias

The 15-30% response rates in HMS raise important questions about selection bias and model fairness. Students experiencing severe mental health crises may be systematically less likely to complete surveys, potentially leading to under-representation of the most vulnerable population. This could affect model calibration in several ways. First, if the training data under-represent students with severe symptoms, these models might be less accurate for precisely those students most in need of intervention. Second, the protective effect of well-being that was observed might be even stronger in the full population if students with very low well-being are under-represented in the sample. Third, socioeconomic factors, including food and housing insecurity, might show different patterns in a more representative sample.

Applying these models outside the U.S. four-year institution context requires careful consideration. Community colleges and minority-serving institutions may serve student populations with different demographic compositions, stress profiles, and mental health trajectories than the predominantly four-year institutions in HMS. Before deployment in these contexts, models should undergo institution-specific validation and potential recalibration. Response bias patterns might also differ across institutional types - for instance, commuter

students at community colleges might have different survey completion rates than residential students at four-year institutions. Despite these limitations, the finding of minimal practical drift within the HMS sample provides encouraging evidence that temporal stability is achievable when monitoring confirms distributional consistency.

D. Well-Being as a Dominant Protective Factor

The striking dominance of well-being in the current models (54.30% of feature importance) carried important messages for intervention design. Traditional mental health approaches emphasize symptom reduction - treating depression, reducing anxiety. These findings suggested a complementary framework: promoting positive mental health [17]. Students with high well-being are dramatically less likely to experience clinically significant depression or anxiety. This has practical implications: universal well-being programs that reach all students may be as important as targeted interventions for high-risk individuals [27].

E. Limitations

Several limitations should be considered. First, the predictor set was constrained by missing data patterns. While the six-predictor model achieved strong performance, the inclusion of additional variables with lower missingness could potentially improve predictive accuracy. Second, the HMS response rates of 15-30% introduced potential selection bias, including possible overrepresentation of female students and individuals more willing to disclose mental health experiences. Although HMS employed stratified sampling procedures, prevalence estimates and model performance may differ in subpopulations less likely to participate in survey research. From a fairness perspective, systematic under-representation of certain demographic groups - such as students of color, first-generation students, or those with lower socioeconomic status - could lead to differential model performance across subgroups. Future work should explicitly audit model performance across demographic levels to identify potential disparities before operational deployment. Third, the three-year observation window captured a specific post-pandemic recovery period. Longer time horizons are needed to determine whether temporal stability persists under different social or institutional conditions [28]. Finally, while HMS includes a diverse set of U.S. institutions, findings may not generalize to non-U.S. educational systems, community colleges, or institutions with substantially different demographic or cultural contexts. True external validation requires testing on independent institutional samples to assess broader generalizability [29].

V. CONCLUSION

Temporal validation of Machine Learning (ML) models for college mental health prediction remains an underexplored prerequisite for real-world deployment. Across three consecutive academic years (2022-2025), this study demonstrated that distributional changes in key predictors were statistically detectable but practically negligible with small effect sizes (all $d < 0.15$) and stable PSI values (< 0.10). Predictive performance remained consistently strong (mean F1 $\approx$ 0.71; AUC $\approx$ 0.80) with minimal temporal degradation (1.8%) under cross-year validation.

These results support the feasibility of deploying static ML-based screening systems in college mental health contexts, contingent on ongoing distributional monitoring. The negligible performance difference between Random Forest (RF) and XGBoost further suggests that robust predictive patterns can be captured without reliance on highly complex modeling approaches, while the lower performance of Logistic Regression (LR) highlights the importance of nonlinear relationships. Well-being emerged as the dominant protective factor, accounting for 54.30% of predictive importance. For researchers, these findings underscore the importance of temporal validation and effect size reporting. For practitioners, the results indicate that ML-based mental health screening is feasible and can be implemented using simple modeling approaches when temporal stability exists.

Future research should focus on longer-term studies spanning five or more years or external validation across diverse institutional contexts, including community colleges, minority-serving institutions, and international universities. Additionally, methodological advances in handling missing data could enable the inclusion of additional predictors that were excluded from the current analysis due to missingness constraints, potentially further improving model performance. Besides these, prospective studies examining whether ML-based screening actually improves student outcomes through earlier intervention would provide crucial evidence for implementation decisions. Finally, research exploring how to operationalize well-being interventions at scale, given its dominant protective role, represents a particularly promising avenue for translating these findings into practice.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this paper.

ACKNOWLEDGMENT

The authors thank the Healthy Minds Network and the research team of the Healthy Minds Study, directed by Dr. Sarah Lipson (Boston University) and Dr. Daniel Eisenberg (UCLA), for providing access to the longitudinal survey data used in this study. This research received no external funding.

DATA AVAILABILITY

The data that support the findings of this study are available from the Healthy Minds Network upon submission of a formal data request and agreement to the applicable data use terms [1]. Requests may be submitted via the Healthy Minds Network data portal. The survey years analyzed in this study are 2022-2023, 2023-2024, and 2024-2025. In accordance with the data use agreement, the raw dataset cannot be publicly redistributed by the authors.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Claude (Anthropic) for language editing and proofreading assistance. The authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] D. Eisenberg, S. K. Lipson, J. Heinze, and S. Zhou, "The Health Minds Study: 2024-2025 Data Report.", Health Minds Network, https://healthymindsnetwork.org/wp-content/uploads/2025/09/2024-2025_HMS-National-Data-Report_Student.pdf.

[2] S. K. Lipson *et al.*, "Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021," *Journal of Affective Disorders*, vol. 306, pp. 138–147, June 2022, https://doi.org/10.1016/j.jad.2022.03.038.

[3] J. Hunt and D. Eisenberg, "Mental Health Problems and Help-Seeking Behavior Among College Students," *Journal of Adolescent Health*, vol. 46, no. 1, pp. 3–10, Jan. 2010, https://doi.org/10.1016/j.jadohealth.2009.08.008.

[4] D. Eisenberg, M. F. Downs, E. Golberstein, and K. Zivin, "Stigma and Help Seeking for Mental Health Among College Students," *Medical Care Research and Review*, vol. 66, no. 5, pp. 522–541, Oct. 2009, https://doi.org/10.1177/1077558709335173.

[5] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, July 2019, https://doi.org/10.1017/S0033291719000151.

[6] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine Learning Approaches for Clinical Psychology and Psychiatry," *Annual Review of Clinical Psychology*, vol. 14, pp. 91–118, May 2018, https://doi.org/10.1146/annurev-clinpsy-032816-045037.

[7] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1-44:37, Nov. 2014, https://doi.org/10.1145/2523813.

[8] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under Concept Drift: A Review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, Dec. 2019, https://doi.org/10.1109/TKDE.2018.2876857.

[9] S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, "Calibration drift in regression and machine learning models for acute kidney injury," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1052–1061, Nov. 2017, https://doi.org/10.1093/jamia/ocx030.

[10] P. C. Austin, D. van Klaveren, Y. Vergouwe, D. Nieboer, D. S. Lee, and E. W. Steyerberg, "Geographic and temporal validity of prediction models: different approaches were useful to examine model performance," *Journal of Clinical Epidemiology*, vol. 79, pp. 76–85, Nov. 2016, https://doi.org/10.1016/j.jclinepi.2016.05.007.

[11] X. Wang, S. Hegde, C. Son, B. Keller, A. Smith, and F. Sasangohar, "Investigating Mental Health of US College Students During the COVID-19 Pandemic: Cross-Sectional Survey Study," *Journal of Medical Internet Research*, vol. 22, no. 9, Sept. 2020, Art. no. e22817, https://doi.org/10.2196/22817.

[12] C. Son, S. Hegde, A. Smith, X. Wang, and F. Sasangohar, "Effects of COVID-19 on College Students' Mental Health in the United States: Interview Survey Study," *Journal of Medical Internet Research*, vol. 22, no. 9, Sept. 2020, Art. no. e21279, https://doi.org/10.2196/21279.

[13] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning," *Clinical Psychological Science*, vol. 5, no. 3, pp. 457–469, May 2017, https://doi.org/10.1177/2167702617691560.

[14] T. A. Burke *et al.*, "Identifying the relative importance of non-suicidal self-injury features in classifying suicidal ideation, plans, and behavior using exploratory data mining," *Psychiatry Research*, vol. 262, pp. 175–183, Apr. 2018, https://doi.org/10.1016/j.psychres.2018.01.045.

[15] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a Brief Depression Severity Measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001, https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

[16] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 2006, https://doi.org/10.1001/archinte.166.10.1092.

[17] E. Diener *et al.*, "New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings," *Social Indicators Research*, vol. 97, no. 2, pp. 143–156, May 2009, https://doi.org/10.1007/s11205-009-9493-y.

[18] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016, https://doi.org/10.1080/00031305.2016.1154108.

[19] F. J. Massey Jr., "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951, https://doi.org/10.1080/01621459.1951.10500769.

[20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Lawrence Erlbaum Associates, 1988.

[21] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Hoboken, NJ, USA: John Wiley & Sons, 2006.

[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, https://doi.org/10.1023/A:1010933404324.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, https://doi.org/10.1145/2939672.2939785.

[24] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Frontiers in Psychology*, vol. 4, Nov. 2013, https://doi.org/10.3389/fpsyg.2013.00863.

[25] G. Widmer and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning*, vol. 23, no. 1, pp. 69–101, Apr. 1996, https://doi.org/10.1023/A:1018046501280.

[26] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, Sept. 2017, https://doi.org/10.1016/j.inffus.2017.02.004.

[27] D. Eisenberg, E. Golberstein, and J. B. Hunt, "Mental Health and Academic Success in College," *The B.E. Journal of Economic Analysis & Policy*, vol. 9, no. 1, Art. no. 40, https://doi.org/10.2202/1935-1682.2191.

[28] K. Zivin, D. Eisenberg, S. E. Gollust, and E. Golberstein, "Persistence of mental health problems and needs in a college student population," *Journal of Affective Disorders*, vol. 117, no. 3, pp. 180–185, Oct. 2009, https://doi.org/10.1016/j.jad.2009.01.001.

[29] S. K. Lipson, A. Kern, D. Eisenberg, and A. M. Breland-Noble, "Mental Health Disparities Among College Students of Color," *Journal of Adolescent Health*, vol. 63, no. 3, pp. 348–356, Sept. 2018, https://doi.org/10.1016/j.jadohealth.2018.04.014.