

Benchmarking Classical Machine Learning Models for College Student Mental Health Screening Using Passive Sensing Data

Prajoko^{a,1*}, Didik Indrayana^{a,2}, and Agung Pambudi^{a,3}

^a Informatics Engineering Study Program, Universitas Muhammadiyah Sukabumi, Sukabumi, West Java, Indonesia

e-mail: Prajoko-ti@ummi.ac.id, ²Didikind@ummi.ac.id, ³Agungpambd@ummi.ac.id

*Corresponding author

Abstract

Mental health risk screening among college students may be supported by passive smartphone sensing; however, model evaluation remains challenging because longitudinal sensing data contain repeated observations within individuals, substantial class imbalance, and incomplete sensor measurements. This study establishes a reproducible benchmark of five classical machine learning algorithms—Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Naive Bayes, and Decision Tree—using the publicly available StudentLife dataset. The analysis comprised 45 independent participants, including five positive-risk and 40 negative-risk participants, represented by 2,716 repeated participant-day **observations** derived from five daily behavioral features. Depression risk was defined at the participant level using the post-semester PHQ-9 threshold of ≥ 10 ; therefore, participant-day records were treated as longitudinal observations rather than independent subjects. Five-fold StratifiedGroupKFold cross-validation was applied using participant identifiers to prevent observations from the same individual from appearing in both training and held-out partitions. Performance was assessed using AUROC, AUPRC, F1-score, precision, and recall, with additional robustness evaluation under 10% and 30% simulated feature corruption. Decision Tree achieved the highest mean AUROC (0.564), F1-score (0.210), and recall (0.480), while Decision Tree and SVM shared the highest mean AUPRC (0.164). Given the small number of independent positive-risk participants and substantial fold-level variability, these findings should be interpreted as preliminary benchmark evidence rather than evidence of model superiority.

Keywords: college student mental health; machine learning; passive sensing; depression risk; participant-independent evaluation; class imbalance

1. Introduction

The psychological well-being of university students is an important public-health concern because it can affect academic achievement, social adjustment, and overall quality of life [1]. The transition to higher education and the demands of campus life are associated with stress, anxiety, and depressive symptoms, creating a need for scalable approaches that can support earlier risk identification [2]. Conventional assessment remains important, but it is often episodic and dependent on retrospective self-report, making it difficult to capture day-to-day fluctuations in students' psychological conditions [7]–[9].

Passive smartphone sensing provides a complementary source of behavioral information by continuously recording patterns such as activity, mobility, social proximity, conversations, and device interaction without requiring repeated manual input [5], [7], [8]. This direction is consistent with the broader development of digital phenotyping and computational mental-health monitoring [6], [9]. Classical machine learning methods such as Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Naive Bayes, and Decision Tree remain useful as transparent and computationally efficient baselines before more complex architectures are introduced [3], [4], [11].

A major methodological concern in passive sensing is information leakage. Longitudinal datasets contain multiple observations from the same participant; therefore, random record-level splitting can place observations from one person in both training and testing sets and produce overly optimistic estimates of generalization. Subject-wise validation strategies such as GroupKFold or Leave-One-Subject-Out are

consequently recommended when the intended use is prediction for unseen participants [14]–[16]. Fair benchmarking also requires the same preprocessing, tuning, and evaluation procedures across all candidate models [13], [17].

Another practical challenge is missing data. Sensor interruptions, hardware limitations, changing phone-use behavior, and non-compliance can produce incomplete passive sensing records [12], [18], [20]. A model that performs well only under complete-data conditions may be unsuitable for continuous monitoring. For this reason, robustness to simulated missingness should be evaluated together with conventional discrimination metrics [13].

The potential of StudentLife and other passive-sensing datasets for examining behavioral patterns associated with student mental health, depression severity, and broader psychopathology has been demonstrated in previous studies [1], [5], [6], [8], [18], [20]. While these studies highlight the value of passive smartphone sensing for mental health research, the longitudinal nature of such data introduces an important methodological challenge because multiple daily observations are repeatedly collected from the same individuals. Consequently, reliable benchmarking requires an evaluation design that explicitly preserves participant independence and prevents information from the same participant from being shared between training and evaluation data. In this context, the novelty of the present study lies in establishing a reproducible and leakage-aware classical machine-learning benchmark that integrates participant-independent partitioning, imbalance-aware evaluation, and robustness assessment under simulated sensing corruption. Rather than seeking to establish the superiority of a particular classifier, this study aims to provide a transparent methodological baseline for evaluating classical machine-learning models for initial depression-risk screening using StudentLife passive-sensing data. The contributions of this study are threefold: (1) a grouped-stratified comparison of five classical machine-learning algorithms under participant-independent validation; (2) an evaluation emphasizing AUPRC and recall in the presence of severe class imbalance; and (3) a robustness assessment under 10% and 30% simulated sensing corruption.

2. Research Methodology

a. Data Collection and Preprocessing

This study used the publicly available StudentLife dataset [5], which contains psychological assessments and longitudinal passive-sensing data collected from college students through smartphones. The original StudentLife cohort consisted of 48 participants. For the present analysis, participants without a usable post-semester PHQ-9 assessment or with insufficient longitudinal sensing coverage were excluded, resulting in a final cohort of 45 participants. Depression risk was defined at the participant level using the post-semester PHQ-9 assessment. Participants with PHQ-9 scores ≥ 10 were categorized as positive risk, whereas those with scores < 10 were categorized as negative risk. The threshold of PHQ-9 ≥ 10 was selected because it is widely used as a screening threshold for clinically relevant depressive symptoms and has established validity for depression screening [21]. The resulting cohort consisted of five positive-risk participants (11.1%) and 40 negative-risk participants (88.9%), indicating substantial class imbalance.

For each participant, the longitudinal passive-sensing data were aggregated into 24-hour intervals. This procedure produced 2,716 participant-day observations from the 45 retained participants, with a median of 64 daily observations per participant and a range of 32–67 days. Each participant-day was represented by five continuous behavioral features: immobile behavior, active behavior, conversation duration, Bluetooth-based proximity, and the longest screen-dark period. Importantly, the 2,716 participant-day records do not represent 2,716 independent individuals or statistically independent samples. Rather, they constitute repeated longitudinal observations nested within 45 independent participants. Thus, the participant represents the independent unit in the study, whereas the participant-day represents the modeling observation used to characterize daily behavioral patterns.

The PHQ-9 outcome was defined at the participant level rather than at the participant-day level. Specifically, the PHQ-9 was not measured daily; instead, each retained participant had a single post-semester PHQ-9-based screening outcome, defined as positive risk for PHQ-9 ≥ 10 and negative risk for PHQ-9 < 10 . For supervised model development, this participant-level outcome was linked through the participant identifier (uid) to all daily sensing observations belonging to the corresponding participant. Consequently, multiple participant-day records from the same individual shared the same participant-level screening label. This labeling procedure should not be interpreted as evidence that depressive symptoms or PHQ-9 scores remained constant across individual study days. Rather, the repeated sensing observations were used as longitudinal behavioral measurements associated with the participant's post-semester PHQ-9-based screening outcome.

Participant identity was explicitly preserved throughout model construction and evaluation to account for the nested structure of the longitudinal observations and to reduce the risk of participant-level information leakage. During cross-validation, all participant-day observations originating from the same individual were assigned exclusively to either the training partition or the held-out partition within a given fold. Consequently, no participant contributed observations to both partitions. Data-dependent preprocessing was performed using only the training partition. Continuous features were standardized using Z-score normalization with parameters estimated from the training data, while missing values were imputed using statistics derived exclusively from the training data. The fitted preprocessing parameters were subsequently applied to the corresponding held-out observations, which remained unseen during preprocessing and model fitting [14]. This participant-independent procedure was designed to ensure that evaluation reflected generalization across participants rather than recognition of participant-specific behavioral patterns.

Table 1. Summary of dataset characteristics

Component	Description
Dataset Name	StudentLife Dataset
Original Cohort	48 participants
Final Eligible Cohort	45 independent participants
Eligibility	Usable post-semester PHQ-9 assessment and sufficient longitudinal passive-sensing coverage
Positive-Risk Participants	5 participants (11.1%)
Negative-Risk Participants	40 participants (88.9%)
Independent Unit	Participant (N = 45)
Outcome Definition	Participant-level depression-risk screening based on post-semester PHQ-9
PHQ-9 Threshold	Positive risk: $\text{PHQ-9} \geq 10$; negative risk: $\text{PHQ-9} < 10$
Temporal Aggregation	24-hour participant-specific intervals
Modeling Observation Unit	Participant-day
Total Daily Observations	2,716 repeated participant-day observations
Daily Observations per Participant	Median: 64 days; range: 32–67 days
Longitudinal Data Structure	Repeated daily observations nested within participants
Number of Behavioral Features	5 continuous daily features
Behavioral Features	Stationary ratio, active ratio, conversation duration, number of unique Bluetooth devices, and longest screen-dark duration
Evaluation Grouping Unit	Participant; all daily observations from the same participant were retained within the same cross-validation fold
Missing-Value Handling	Imputation parameters estimated exclusively from the training fold and subsequently applied to held-out observations
Feature Scaling	Z-score standardization using parameters estimated exclusively from the training fold

Although the passive-sensing predictors were represented as repeated participant-day observations, the PHQ-9 outcome was defined at the participant level rather than at the daily observation level. Each retained participant was assigned a single post-semester PHQ-9-based screening outcome—positive risk for $\text{PHQ-9} \geq 10$ or negative risk for $\text{PHQ-9} < 10$ —which was linked through the participant identifier (uid) to all corresponding daily sensing observations for supervised model development. Consequently, daily records from the same participant shared the same participant-level screening label. Importantly, these

repeated participant-day records should not be interpreted as independent samples; the effective independent cohort consisted of 45 participants, including only five positive-risk participants.

This labelling approach should not be taken to mean that depression symptoms or PHQ-9 scores were seen or presumed to be the same on each research day. Instead, the participant's overall PHQ-9-based screening outcome was estimated using the daily sensing records as repeated behavioural assessments. As a result, rather than 2,716 separate samples, the 2,716 participant-day records are longitudinal repeated observations nested within 45 persons. All daily observations from a particular participant were kept together within the same cross-validation fold to account for this dependence and avoid participant-level information leakage. This ensured that no person provided observations to both the training and held-out test sets.

b. Benchmark Pipeline and Candidate Models

Figure 1 summarizes the computational pipeline. The procedure begins with StudentLife data preparation, followed by subject-wise splitting, training of the candidate single-model algorithms, uniform hyperparameter tuning within training data, performance measurement, and missing-data robustness testing. Model recommendation is based on both screening performance and stability rather than on a single metric.

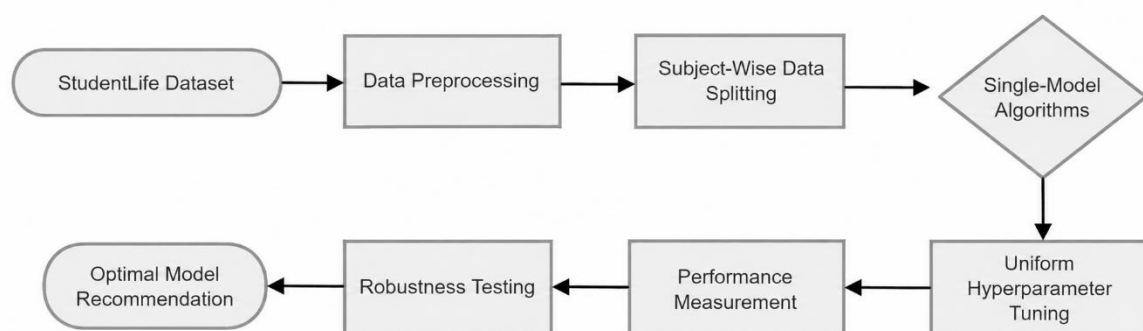


Figure 1. Participant-Independent Benchmarking Pipeline

The benchmark is deliberately restricted to five classical single-model algorithms: Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naive Bayes (NB), and Decision Tree (DT). These models represent different inductive assumptions while remaining relatively simple to reproduce and interpret. Ensemble and deep learning methods were excluded from this initial benchmark so that the comparison could establish a clear conventional baseline.

Table 2. Candidate algorithms and hyperparameter tuning plan

Model	Fixed Configuration	Hyperparameter Search Space	Benchmark Role
Logistic Regression	class_weight='balanced', max_iter=1000, random_state=42	C = {0.1, 1.0, 10.0}; penalty = {L2}; solver = {lbfgs, liblinear}	Linear screening baseline
Support Vector Machine	probability=True, class_weight='balanced', random_state=42	C = {0.1, 1.0, 10.0}; kernel = {linear, RBF}; gamma = {scale, auto}	Margin-based comparator
k-Nearest Neighbors	KNeighborsClassifier()	n_neighbors = {3, 5, 7}; weights = {uniform, distance};	Non-parametric comparator

Model	Fixed Configuration	Hyperparameter Search Space	Benchmark Role
		metric = {euclidean, manhattan}	
Naive Bayes	GaussianNB()	var_smoothing = { 10^{-9} , 10^{-8} , 10^{-7} }	Lightweight probabilistic baseline
Decision Tree	class_weight='balanced', random_state=42	max_depth = {None, 5, 10, 15}; min_samples_split = {2, 5, 10}; min_samples_leaf = {1, 2, 4}	Interpretable non-linear baseline

c. Evaluation Protocol and Performance Measures

All candidate models were evaluated using five-fold grouped-stratified cross-validation implemented with StratifiedGroupKFold. The participant identifier (uid) was specified as the grouping variable so that all participant-day observations belonging to the same individual were assigned exclusively to either the training or held-out partition within each fold. This participant-independent partitioning prevented observations from the same participant from appearing simultaneously in the training and evaluation data, thereby reducing the risk of cross-subject information leakage [14], [16]. Stratification was additionally applied to maintain representation of the minority positive-risk class across the five validation folds.

The final cohort consisted of 45 independent participants, including five positive-risk and 40 negative-risk participants. Under the five-fold grouped-stratified scheme, each held-out fold contained nine participants, comprising one positive-risk and eight negative-risk participants, while the corresponding training partition contained 36 participants, comprising four positive-risk and 32 negative-risk participants. Because the number of valid sensing days varied among participants, the number of participant-day observations in the held-out partitions ranged from 527 to 552 observations. The resulting participant-level class distribution across the five validation folds is summarized in Table 3. Importantly, the 2,716 participant-day observations represent repeated longitudinal measurements nested within the 45 participants and should not be interpreted as 2,716 statistically independent samples.

Table 3. Participant-Level Class Distribution Across the Five Grouped-Stratified Validation Folds

Fold	Training N	Pos	Neg	Held-Out N	Pos	Neg	Participant-Days
1	36	4	32	9	1	8	527
2	36	4	32	9	1	8	536
3	36	4	32	9	1	8	551
4	36	4	32	9	1	8	552
5	36	4	32	9	1	8	550

Model development and performance assessment were conducted at the participant-day level because the input representation consisted of daily behavioral feature vectors. Each participant-day record was linked to the PHQ-9-based screening outcome of its corresponding participant. Nevertheless, participant independence was strictly preserved during data partitioning, as all daily observations from a given participant were assigned exclusively to either the training or held-out partition within each fold. Therefore, participant-independent evaluation in this study refers to generalization across unseen participants through subject-wise partitioning rather than to aggregation of daily predictions into a single participant-level prediction.

Accordingly, performance metrics were calculated from held-out participant-day predictions. The reported results should therefore be interpreted as the performance of the candidate models in discriminating repeated daily behavioral observations associated with participant-level depression-risk status among unseen participants, rather than as direct estimates of participant-level screening performance.

Participant-level aggregation was not included in the present benchmark. This distinction is particularly important because the effective independent cohort consisted of only 45 participants, including five positive-risk participants, with only one positive-risk participant represented in each held-out fold. Consequently, the participant-day results are treated as preliminary benchmark evidence and are not interpreted as evidence based on 2,716 independent individuals.

Hyperparameter tuning was performed using GridSearchCV with the same five-fold StratifiedGroupKFold scheme and participant identifier (uid) supplied as the grouping variable. For each candidate model, the complete hyperparameter search space is reported in Table 2. The modeling pipeline consisted of StandardScaler, SMOTE (random_state=42), and the corresponding classifier, ensuring that feature scaling and minority-class oversampling were fitted within the training data. SMOTE was applied only as part of the model-training procedure to address the substantial class imbalance and was not applied to held-out evaluation observations. Multiple evaluation metrics were computed during grid search, including AUROC, AUPRC, F1-score, precision, and recall, while average precision (AUPRC) was specified as the refitting criterion for selecting the preferred hyperparameter configuration. Random state 42 was used for Logistic Regression, SVM, Decision Tree, and SMOTE where applicable.

Model discrimination was assessed using the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC), together with F1-score, precision, and recall (sensitivity). Because the positive-risk group was substantially underrepresented, AUPRC and recall were emphasized when interpreting screening-related performance. Results were summarized as the mean $\pm$ standard deviation across the five validation folds. Given that each held-out fold contained only one positive-risk participant, fold-level minority-class metrics are sensitive to the observations contributed by a single positive participant. Therefore, differences among candidate models were interpreted cautiously and regarded as preliminary comparative benchmark evidence rather than definitive evidence of model superiority.

Robustness to incomplete sensing was evaluated by simulating random feature corruption at two levels, 10% and 30%. At each corruption level, feature cells were randomly selected and the selected values were replaced with zero to represent interrupted or unavailable sensor readings. A fixed random seed of 42 was used to ensure reproducibility. The same five-fold StratifiedGroupKFold procedure was subsequently applied to the corrupted feature matrices, with AUPRC serving as the primary robustness metric. Changes in AUPRC relative to the baseline condition were reported as Δ AUPRC. This procedure was intended as a controlled sensitivity analysis of model behavior under simulated sensing corruption and should not be interpreted as representing all forms of missingness that may occur in real-world passive-sensing data.

3. Results and Discussion

a. Predictive Performance

Table 4 presents the predictive performance of the five candidate models under participant-independent evaluation. The reported metrics were calculated from held-out participant-day observations, while participant identity was strictly preserved during data partitioning so that all daily observations from the same participant remained exclusively within either the training or held-out partition. Therefore, the results should be interpreted as predictive performance on repeated daily behavioral observations from unseen participants rather than as direct participant-level screening performance.

Decision Tree achieved the highest mean AUROC (0.564), F1-score (0.210), and recall (0.480) among the evaluated models. SVM and Decision Tree obtained the same mean AUPRC (0.164), while SVM showed slightly higher mean precision than Decision Tree (0.140 versus 0.136). Logistic Regression achieved the second-highest mean recall (0.417), although its mean AUROC and AUPRC were lower. In contrast, Naive Bayes produced the lowest mean recall (0.066) and F1-score (0.042) among the five candidate models. These numerical differences describe the observed mean performance under the present evaluation setting and should not be interpreted as evidence of definitive superiority of one algorithm over the others.

Table 4. Performance of single-model algorithms on StudentLife using subject-wise evaluation

Model	AUROC	AUPRC	F1	Precision	Recall
Dummy Classifier	0.505 $\pm$ 0.025	0.117 $\pm$ 0.009	0.135 $\pm$ 0.040	0.122 $\pm$ 0.036	0.151 $\pm$ 0.044
LR	0.492 $\pm$ 0.138	0.139 $\pm$ 0.041	0.174 $\pm$ 0.074	0.111 $\pm$ 0.046	0.417 $\pm$ 0.197

Model	AUROC	AUPRC	F1	Precision	Recall
SVM	0.518 $\pm$ 0.174	0.164 $\pm$ 0.069	0.193 $\pm$ 0.093	0.140 $\pm$ 0.075	0.341 $\pm$ 0.139
k-NN	0.508 $\pm$ 0.107	0.128 $\pm$ 0.034	0.168 $\pm$ 0.058	0.115 $\pm$ 0.042	0.317 $\pm$ 0.098
NB	0.473 $\pm$ 0.170	0.121 $\pm$ 0.049	0.042 $\pm$ 0.021	0.041 $\pm$ 0.014	0.066 $\pm$ 0.067
DT	0.564 $\pm$ 0.144	0.164 $\pm$ 0.067	0.210 $\pm$ 0.087	0.136 $\pm$ 0.058	0.480 $\pm$ 0.179

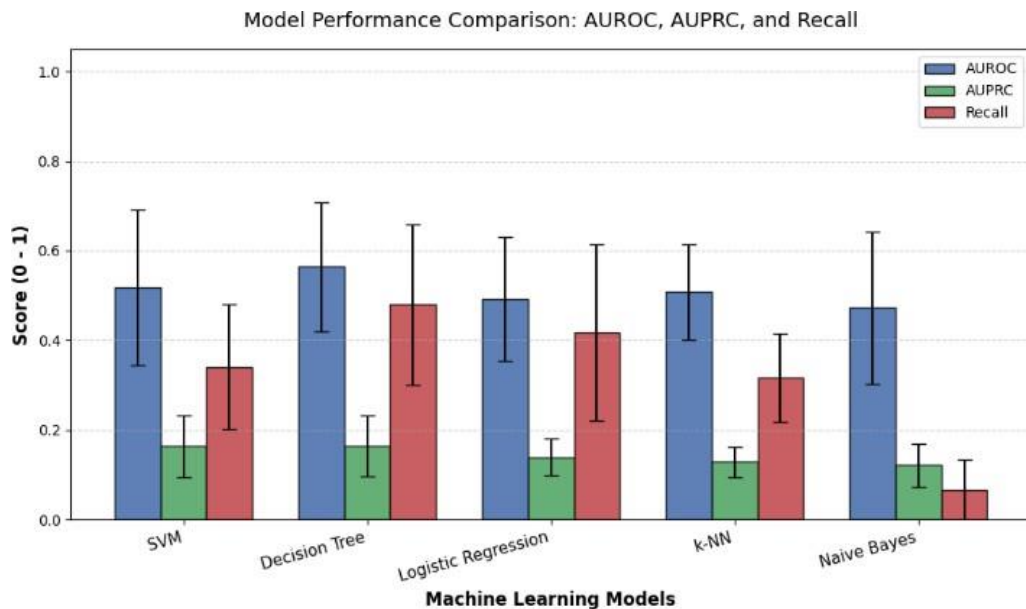


Figure 2. Comparison of AUROC, AUPRC, and recall across the candidate models.

The observed performance differences should be interpreted cautiously in relation to both class imbalance and the limited number of independent participants. Although Decision Tree produced the highest mean recall, F1-score, and AUROC and shared the highest mean AUPRC with SVM, substantial fold-level variability was observed. For example, Decision Tree obtained an AUROC of 0.564 ± 0.144 , whereas SVM obtained 0.518 ± 0.174 . Their AUPRC values were also identical in terms of the mean (0.164), with standard deviations of 0.067 and 0.069, respectively. The overlap in variability indicates that differences in mean performance alone provide insufficient evidence to establish a consistently superior classifier.

The absolute AUPRC values were also modest relative to the positive-class prevalence of approximately 0.111. The highest mean AUPRC of 0.164 therefore reflects only a limited improvement over the prevalence level. At the same time, models with comparatively higher recall did not demonstrate correspondingly high precision. Decision Tree, for example, achieved a mean recall of 0.480 with a mean precision of 0.136, while Logistic Regression achieved a mean recall of 0.417 with a mean precision of 0.111. This pattern illustrates the difficulty of identifying positive-risk observations under substantial class imbalance and supports the use of multiple complementary performance measures rather than relying on a single metric.

An additional consideration is that the reported metrics were derived from repeated participant-day observations nested within only 45 independent participants, including five positive-risk participants. Under the five-fold grouped-stratified evaluation, each held-out fold contained only one positive-risk participant. Consequently, the performance observed within an individual fold may be sensitive to the behavioral observations contributed by that participant, particularly for minority-class measures such as recall, precision, F1-score, and AUPRC. The present results should therefore be regarded as preliminary participant-independent benchmark evidence at the participant-day level rather than as direct estimates of participant-level screening performance.

Taken together, the results indicate that the evaluated classical machine-learning models exhibit different trade-offs between discrimination, precision, and sensitivity under the present benchmark setting. Decision Tree showed comparatively higher mean recall and F1-score, whereas SVM provided a similar mean AUPRC with slightly higher precision. However, the combination of substantial class imbalance,

fold-level variability, and the small number of independent positive-risk participants limits the strength of conclusions that can be drawn from these numerical differences. The findings are therefore more appropriately interpreted as a comparative baseline for subsequent research than as evidence of definitive algorithmic superiority [13], [19].

b. Robustness Under Simulated Missingness

The robustness analysis examined changes in model performance under controlled levels of simulated feature corruption. Specifically, 10% and 30% of feature values were randomly selected and replaced with zero, following the procedure described in the evaluation protocol. Mean AUPRC was used to compare model performance between the baseline and simulated corruption conditions, while Δ AUPRC represented the change relative to the corresponding baseline performance.

SVM and Decision Tree showed decreases in mean AUPRC under both simulated missingness levels. SVM decreased from a baseline AUPRC of 0.164 to 0.148 at 10% missingness (Δ AUPRC = -0.017) and to 0.157 at 30% missingness (Δ AUPRC = -0.008). Decision Tree decreased from 0.164 to 0.137 at 10% missingness (Δ AUPRC = -0.026) and to 0.154 at 30% missingness (Δ AUPRC = -0.009). In comparison, k-NN showed only small changes, increasing from 0.128 to 0.133 at 10% missingness before decreasing to 0.123 at 30% missingness.

Table 5. Robustness analysis under simulated missingness scenarios

Model	Baseline AUPRC	10% Missing	Δ 10%	30% Missing	Δ 30%
LR	0.139	0.143	+0.004	0.149	+0.010
SVM	0.164	0.148	-0.017	0.157	-0.008
k-NN	0.128	0.133	+0.005	0.123	-0.005
NB	0.121	0.133	+0.012	0.129	+0.008
DT	0.164	0.137	-0.026	0.154	-0.009

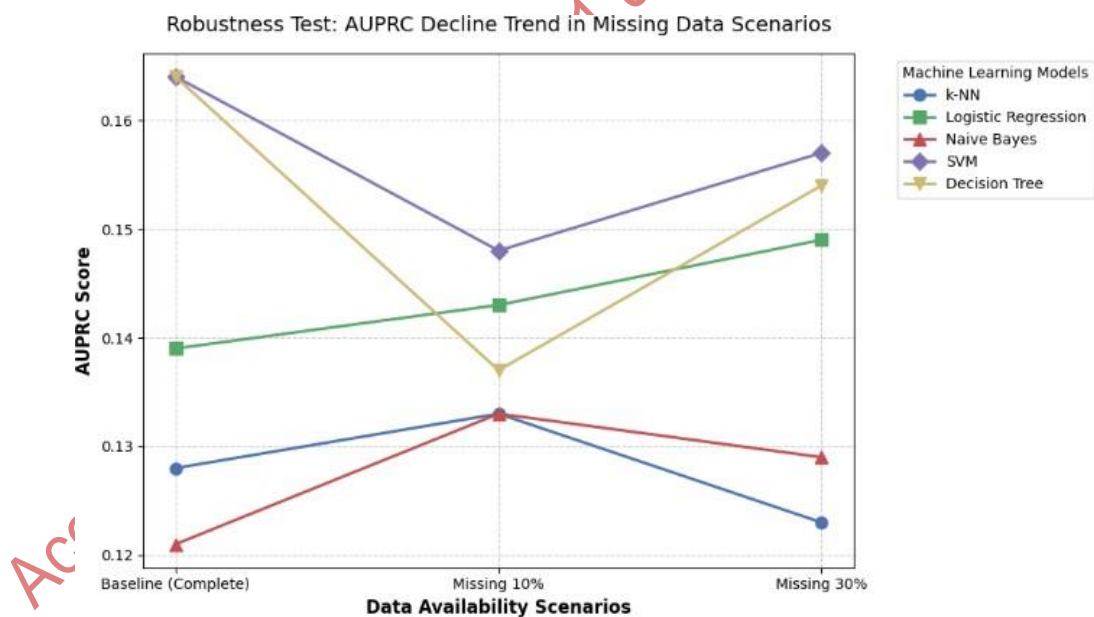


Figure 3. AUPRC trends under complete, 10% missing, and 30% missing feature conditions.

Logistic Regression and Naive Bayes showed small positive changes in mean AUPRC under the simulated missingness conditions. Logistic Regression increased from 0.139 at baseline to 0.143 and 0.149 at 10% and 30% missingness, respectively, whereas Naive Bayes increased from 0.121 to 0.133 and 0.129. These small increases should not be interpreted as evidence that missing data improve predictive performance. Rather, together with the small decreases observed for the other models, they indicate that performance estimates may vary under the particular masking conditions applied to a relatively small and imbalanced dataset.

The results also show that changes in AUPRC were not consistently proportional to the level of simulated missingness. For example, SVM experienced a larger reduction at 10% missingness than at 30%,

while Logistic Regression showed a small increase at both levels. Therefore, the observed Δ AUPRC values are more appropriately interpreted as indicators of model sensitivity under the specific simulated conditions than as evidence of a systematic relationship between the amount of missing data and predictive performance.

The missingness analysis reinforces the importance of considering performance stability when evaluating passive-sensing models. Real-world sensing records can be interrupted by technical limitations and user behavior [12], [18], [20]. However, the present experiment represents a controlled simulation of missingness and does not reproduce the full complexity of missing-data patterns that may occur in naturalistic passive-sensing settings. Accordingly, the results should be interpreted as a sensitivity analysis under the specified simulated conditions rather than as a comprehensive assessment of robustness to all forms of real-world sensing incompleteness.

Overall, no candidate model demonstrated a uniformly dominant combination of baseline discrimination and stability across the two simulated missingness conditions. The relatively small and non-monotonic changes in AUPRC further suggest that robustness differences among the candidate models should be interpreted cautiously. Within the scope of the present benchmark, these findings provide preliminary evidence of how the evaluated classical machine-learning models respond to simulated sensing incompleteness, while more comprehensive evaluation of naturally occurring missingness remains an important direction for future research.

c. Implications and Limitations

The findings of this study indicate that the evaluated classical machine-learning models exhibit different predictive characteristics rather than a consistently dominant performance pattern. Decision Tree showed a comparatively recall-oriented profile, whereas SVM achieved the same mean AUPRC with slightly higher precision. These differences provide a useful empirical baseline for understanding model behavior under severe class imbalance and longitudinal passive-sensing conditions. Nevertheless, the benchmark is not intended to function as a diagnostic system or to establish the superiority of a particular classifier. Its methodological value lies primarily in providing a transparent comparative evaluation in which participant identity is preserved throughout model assessment. The use of subject-wise splitting is an important methodological safeguard because it evaluates generalization to participants who were not seen during training [14], [16].

A central limitation concerns the distinction between modeling observations and statistically independent participants. Although the dataset contains 2,716 participant-day observations, these observations are repeated longitudinal measurements nested within only 45 participants. The positive-risk class is particularly limited, consisting of five independent participants. Consequently, each held-out fold contains only one positive-risk participant under the five-fold grouped-stratified evaluation. Performance within an individual fold may therefore be influenced by the behavioral observations contributed by that participant. For this reason, the 2,716 participant-day records should not be interpreted as an equivalent number of independent samples. The effective independent cohort remains 45 participants, and the small number of positive-risk participants necessarily limits the stability and generalizability of the resulting performance estimates.

Another limitation relates to the level at which predictive performance was quantified. Participant-independent partitioning ensured that daily observations belonging to the same individual were not distributed across both training and held-out data within a given fold. However, AUROC, AUPRC, F1-score, precision, and recall were calculated from held-out participant-day predictions rather than from predictions aggregated into a single estimate for each participant. Participant-independent evaluation should therefore be distinguished from direct participant-level performance evaluation. The reported metrics characterize the ability of the models to discriminate repeated daily behavioral observations associated with participant-level risk status among unseen participants. Accordingly, they should not be interpreted as direct estimates of participant-level screening accuracy, particularly given the limited number of independent positive-risk participants available in the present cohort.

The temporal relationship between the predictors and outcome also defines an important boundary for interpretation. Passive-sensing predictors were represented by longitudinal participant-day behavioral observations, whereas the PHQ-9 outcome was defined as a single post-semester screening status for each participant. Thus, individual daily sensing records were associated with a participant-level outcome rather than with independently measured daily PHQ-9 scores. The resulting predictions should not be interpreted as estimates of day-specific depressive states or as evidence of daily changes in depressive symptoms.

Instead, the present benchmark examines associations between longitudinal behavioral observations and the participant's post-semester PHQ-9-based risk status. This distinction is important when considering how the reported predictive performance relates to the intended role of passive sensing as a preliminary screening approach.

Several additional factors further constrain the scope of the conclusions. The analysis was conducted using a single public dataset, five behavioral features, and one binary PHQ-9 threshold. Model performance may therefore differ when evaluated in other student populations, with alternative target definitions, broader behavioral representations, or different temporal aggregation strategies. In addition, the robustness assessment considered simulated missingness at only two levels and cannot represent the diversity or temporal structure of missing-data mechanisms encountered in naturalistic passive-sensing environments. These limitations, together with the observed fold-level variability and the small independent cohort, support a cautious interpretation of the findings. The reported results are more appropriately viewed as preliminary benchmark evidence than as conclusive evidence of predictive effectiveness or algorithmic superiority.

Future extensions can build on this baseline by evaluating richer temporal representations, sequence models, explainability techniques, and privacy-preserving deployment strategies [10]–[12]. Further investigations should also consider participant-level aggregation of daily predictions, alternative label definitions, repeated or nested subject-wise validation, additional missingness mechanisms, and external validation using larger and more diverse student cohorts. Such analyses would help determine whether the behavioral patterns observed in the present study remain stable across individuals, temporal settings, and populations. They would also provide stronger evidence regarding participant-level screening performance and model robustness under realistic sensing conditions. These methodological extensions are particularly important before the benchmark is translated from an experimental comparison into any form of campus-level screening application.

4. Conclusion

Using passive smartphone sensing data, this study establishes a reproducible classical machine-learning benchmark for depression-risk screening research among college students. The analysis involved 45 independent participants, including five positive-risk participants, represented by 2,716 repeated participant-day observations. A consistent modeling pipeline and five-fold grouped-stratified evaluation using StratifiedGroupKFold were employed to reduce cross-subject information leakage while preserving representation of the minority positive-risk class across validation folds. Decision Tree achieved the highest mean AUROC, F1-score, and recall, while Decision Tree and SVM shared the highest mean AUPRC. However, the moderate absolute performance, substantial fold-level variability, and limited number of independent positive-risk participants do not support a definitive claim of model superiority.

The main contribution of this study is therefore a transparent and reproducible benchmark integrating participant-independent partitioning, imbalance-aware evaluation, and robustness assessment under simulated sensing corruption. Because the reported performance metrics were calculated from participant-day predictions, the findings should be interpreted as preliminary participant-independent benchmark evidence rather than as direct estimates of participant-level screening performance. Furthermore, the longitudinal sensing observations were associated with a single post-semester PHQ-9-based outcome and should not be interpreted as predictions of day-specific depressive states. Future studies should validate these findings using larger and independent student cohorts, participant-level prediction aggregation, broader behavioral representations, additional missing-data mechanisms, and external validation before practical screening deployment is considered.

References

- [1] M. M. Qirtas, E. Zafeiridi, E. B. White, and D. Pesch, "The relationship between loneliness and depression among college students: Mining data derived from passive sensing," *Digit. Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231211104.
- [2] W. Wang et al., "First-Gen Lens: Assessing mental health of first-generation students across their first year at college using mobile sensing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, Jul. 2022, doi: 10.1145/3543194.
- [3] N. N. I. Moubarak, N. M. M. Omar, and V. N. Youssef, "Smartphone-sensor-based human activities classification for forensics: a machine learning approach," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 33, Aug. 2024, doi: 10.1186/s43067-024-00157-x.

- [4] M. Ehsan, "Comparison of the Predictive Models of Human Activity Recognition (HAR) in Smartphones," *UMT Artificial Intelligence Review*, vol. 1, no. 2, pp. 27–35, Dec. 2021, doi: 10.32350/AIR.0102.03.
- [5] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Association for Computing Machinery, Inc., 2014, pp. 3–14, doi: 10.1145/2632048.2632054.
- [6] S. Akbarova et al., "Improving Depression Severity Prediction from Passive Sensing: Symptom-Profiling Approach," *Sensors (Basel)*, vol. 23, no. 21, Oct. 2023, doi: 10.3390/s23218866.
- [7] L. MacLeod et al., "A mobile sensing app to monitor youth mental health: Observational pilot study," *JMIR Mhealth Uhealth*, vol. 9, no. 10, Oct. 2021, doi: 10.2196/20638.
- [8] Y. Terhorst, E. M. Messner, K. O. Asare, C. Montag, C. Kannen, and H. Baumeister, "Investigating Smartphone-Based Sensing Features for Depression Severity Prediction: Observation Study," *J. Med. Internet Res.*, vol. 27, 2025, doi: 10.2196/55308.
- [9] Y. Zhang et al., "Large-scale digital phenotyping: Identifying depression and anxiety indicators in a general UK population with over 10,000 participants," *J. Affect. Disord.*, vol. 375, pp. 412–422, Apr. 2025, doi: 10.1016/j.jad.2025.01.124.
- [10] E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, "Toward explainable AI (XAI) for mental health detection based on language behavior," *Front Psychiatry*, vol. 14, Dec. 2023, doi: 10.3389/fpsyt.2023.1219479.
- [11] V. Quang Tran and H. Byeon, "Explainable hybrid tabular Variational Autoencoder and feature Tokenizer Transformer for depression prediction," *Expert Syst. Appl.*, vol. 265, Mar. 2025, doi: 10.1016/j.eswa.2024.126084.
- [12] B. Zou et al., "Sequence Modeling of Passive Sensing Data for Treatment Response Prediction in Major Depressive Disorder," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1786–1795, 2023, doi: 10.1109/TNSRE.2023.3260301.
- [13] A. Perez-Lebel, G. Varoquaux, M. Le Morvan, J. Josse, and J.-B. Poline, "Benchmarking missing-values approaches for predictive models on health databases," Feb. 2022. [Online]. Available: <http://arxiv.org/abs/2202.10580>.
- [14] S. U. Rehman, A. Ali, A. M. Khan, and C. Okpala, "Human Activity Recognition: A Comparative Study of Validation Methods and Impact of Feature Extraction in Wearable Sensors," *Algorithms*, vol. 17, no. 12, Dec. 2024, doi: 10.3390/a17120556.
- [15] M. F. Tørring, A. Logacjov, S. M. Brændvik, A. Ustad, K. Roeleveld, and E. M. Bardal, "Validation of two novel human activity recognition models for typically developing children and children with Cerebral Palsy," *PLoS One*, vol. 19, no. 9, Sep. 2024, doi: 10.1371/journal.pone.0308853.
- [16] F. M. Calatrava-Nicolás, S. Miyauchi, V. F. Rey, P. Lukowicz, T. Stoyanov, and O. M. Mozos, "Embedded Inter-Subject Variability in Adversarial Learning for Inertial Sensor-Based Human Activity Recognition," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE Computer Society, 2025, doi: 10.1109/MLSP62443.2025.11204225.
- [17] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *J. Biomed. Inform.*, vol. 113, Jan. 2021, doi: 10.1016/j.jbi.2020.103621.
- [18] B. Lamichhane, N. Moukaddam, and A. Sabharwal, "Mobile sensing-based depression severity assessment in participants with heterogeneous mental health conditions," *Sci. Rep.*, vol. 14, no. 1, p. 18808, Aug. 2024, doi: 10.1038/s41598-024-69739-z.
- [19] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS One*, vol. 17, no. 1, Jan. 2022, doi: 10.1371/journal.pone.0262838.
- [20] W. R. Ringwald, G. King, C. E. Vize, and A. G. C. Wright, "Passive Smartphone Sensors for Detecting Psychopathology," *JAMA Netw. Open*, vol. 8, no. 7, Jul. 2025, doi: 10.1001/jamanetworkopen.2025.19047.
- [21] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.