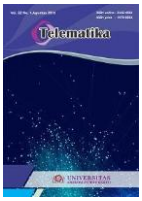


Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemdiktisaintek, No.10/C/C3/DT.05.00/2025

Random Forest Based Prediction of Student Stress Levels from Digital Activity Data

Alphin Stephanus[1], Sri W. Ginting[2], Junus J. S. Kufla[3], Syukri G. Suatkab[4], Thenny D. Salamoni[5]

[1.2, 3, 4, 5] Teknik Informatika, Jurusan Teknik Elektro, Politeknik Negeri Ambon, Ambon, Indonesia

ARTICLE INFO

History of the article:
Received May 20, 2026
Revised July 6, 2026
Accepted August 12, 2026

Keywords:
Student stress
Digital activity
Random forest
DASS-21
Streamlit

Correspondence:

E-mail:
stephanusalphin@gmail.com

ABSTRACT

Early-semester students experience academic adaptation while relying intensively on digital devices, creating a need for accessible, non clinical stress screening. This study aimed to determine whether six self-reported digital-activity variables could distinguish low, moderate, and high DASS-21 derived stress categories and support a student facing screening prototype. A quantitative cross sectional survey and software prototyping design involved 66 Informatics Engineering students. The analytical procedure combined median imputation, standardization, SMOTE within each training fold, Random Forest classification, repeated stratified five fold cross validation with ten repeats, comparison algorithms, class-specific metrics, permutation importance, learning-curve analysis, and functional testing. Metric auditing produced a single out of fold accuracy of 62.12% and macro F1-score of 53.88%. Under repeated validation, the SMOTE Random Forest pipeline achieved 61.97 ± 10.92% accuracy and 55.82 ± 9.87% macro F1-score. The majority DummyClassifier obtained higher accuracy at 65.16 ± 3.47% but only 26.29 ± 0.84% macro F1-score and zero recall for moderate and high stress. SMOTE Random Forest recall was 20.67 ± 23.94% for moderate stress and 80.00 ± 30.00% for high stress. Logistic Regression produced the highest comparison-model accuracy of 70.89%, although its macro F1-score of 55.23% was slightly below that of SMOTE Random Forest. Total screen time was the only predictor with clearly positive permutation importance. Overlapping feature distributions, unstable minority class estimates, and a persistent training validation gap limited performance. StresCheck therefore constitutes an exploratory proof of concept and requires larger, balanced, externally validated data before screening deployment.

1. INTRODUCTION

University students face academic workload, social adjustment, and changes in daily routines, especially during transition into higher education. A Southeast Asian systematic review reported a substantial burden of mental health problems and emphasized locally appropriate early detection (Dessauvagie et al., 2022). High perceived stress has also been documented during technology mediated university learning (Malik & Javed, 2021). Academic burnout is associated with psychological distress, sleep quality, and lifestyle factors (Pereira et al., 2025).

Digital technology is central to learning, communication, and entertainment, yet prolonged or poorly regulated use may contribute to technostress. Screen exposure has been associated with adverse psychological outcomes (Boers et al., 2019; Twenge & Campbell, 2018), while Indonesian studies report relationships between intensive social media use and student mental health (Hilda & Roy Purwanto, 2024; Sa'diyah et al., 2022). Hybrid learning may amplify technostress when institutional and technological support is insufficient (Daud, 2025).

Stress screening commonly uses self report instruments. Machine learning can complement periodic assessment by learning nonlinear relationships among behavioral variables and supporting classification across heterogeneous data (Tufail et al., 2023). Random Forest and other supervised algorithms have identified mental health or burnout risk profiles in international and Indonesian contexts (Jamilah, 2026; Ku & Min, 2024; Pereira et al., 2025; Sharif et al., 2023). However, student mental health models require transparent validation, privacy, fairness, and careful interpretation because prediction is not clinical diagnosis (Drira et al., 2024).

Random Forest combines decision trees, reduces the variance of a single tree, handles heterogeneous tabular data, and yields feature importance estimates. Recent studies published in Telematika have demonstrated the relevance of Random Forest within comparative classification and ensemble evaluation, including the effects of feature selection and model optimization (Defit et al., 2024), as well as cross validated assessment using accuracy, precision, recall, F1-score, AUC-ROC, and feature importance (Beny, 2026). Stratified cross validation is valuable when class frequencies differ, while class level precision, recall, and F1-score disclose behavior hidden by overall accuracy (Yang & Berdine, 2024).

Recent evidence also exposes a specific methodological gap. Reviews of machine learning studies on anxiety and stress in college students report wide variation in predictors, algorithms, and validation practices, with Logistic Regression and Support Vector Machine frequently competitive (Daza et al., 2023). Large survey studies and passive sensing investigations use richer demographic, behavioral, mobility, or physiological data (Ratul et al., 2023; Shvetcov et al., 2024), whereas cross dataset wearable research shows that models trained on small, single protocol samples may generalize poorly (Vos et al., 2023). Prospective digital phenotyping work likewise emphasizes validation across cohorts rather than performance in one development sample (Currey & Torous, 2022). Few studies therefore examine whether a minimal set of routinely accessible digital activity variables can support three class stress screening in a local vocational college cohort while simultaneously reporting a majority baseline, class imbalance experiments, repeated validation uncertainty, class specific errors, and model agnostic feature importance.

The research problem was formulated as follows: can six accessible digital activity variables distinguish low, moderate, and high DASS-21-derived stress categories more informatively than a majority class rule, and which errors limit their use for early educational screening. The proposed solution is StresCheck, an exploratory pipeline that combines leakage controlled preprocessing, imbalance aware Random Forest modeling, comparative algorithms, class specific evaluation, and a Streamlit interface. The objective was to evaluate predictive performance, stability, minority class sensitivity, and feature contribution, and then assess whether the model could be implemented as a technically functional prototype. Its novelty lies not in a new algorithm but in the integration of a data minimal predictor set, transparent repeated validation, explicit intermediate class error analysis, and a student facing workflow in an under studied Indonesian vocational college setting. Scientifically, the study tests the limits of low burden digital behavior variables rather than assuming that a deployable model follows from overall accuracy alone.

2. RESEARCH METHODS

2.1. Research Design, Materials, And Data Collection

A quantitative survey and software prototyping design was applied at Politeknik Negeri Ambon from March to June 2026. The dataset contained 66 early semester students. Digital activity responses referred to built in mobile statistics such as Digital Wellbeing or Screen Time. The stress component of DASS-21 supplied low, moderate, and high target classes; consequently, the output is a questionnaire linked screening result rather than a medical diagnosis.

The 66 participants constituted a bounded exploratory cohort; no a priori machine learning sample size calculation was performed. The learning curve analysis was therefore used diagnostically rather than to claim a minimum sufficient sample. Training macro F1 remained 1.00 while validation macro F1 stayed approximately 0.54–0.62 across the evaluated training sizes, indicating substantial variance and limited generalizability. The eight observation high stress class further widened uncertainty.

Formal ethical approval and documented informed consent for research publication were not obtained before data collection. The data originated from a minimal risk, non interventional questionnaire administered to students within the authors' academic program. The analytical dataset was handled without direct personal identifiers, access was limited to the research team, and results are presented only at aggregate level. Questionnaire responses and model predictions were not used for clinical diagnosis or to initiate automated academic or administrative decisions. The absence of prospective ethical review is acknowledged explicitly as a procedural limitation.

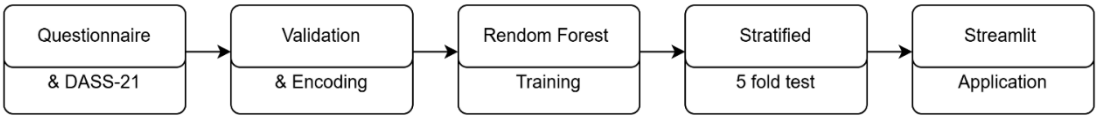


Figure 1. Research and system development workflow.

2.2. Variables and processing

Logical validation required social media, academic, and entertainment durations not to exceed total screen time. Numerical and ordinal inputs were prepared for Random Forest training. The variables are summarized in Table 1. Spearman analysis found no strong pairwise association among the six predictors using the prespecified $|\rho| \geq 0.70$ threshold. Nevertheless, sensitivity analysis showed that predictor composition materially affected performance: the six feature model achieved macro F1 = 0.5582 ± 0.0987; removing total screen time reduced it to 0.4198 ± 0.1378; and retaining total screen time while removing the three component-duration variables increased it to 0.6873 ± 0.1012. Thus, screen time carried substantial predictive information, while the component durations may have introduced redundancy or noise despite the absence of pairwise $|\rho| \geq 0.70$.

The research materials comprised the anonymized 66 row tabular dataset, the six digital activity predictors, the DASS-21 derived target labels already recorded in the dataset, Python libraries, and the Streamlit prototype. Each row was treated as one independent participant observation. Duration variables were assumed to refer to the same daily reporting period and to be consistent with the device statistics

consulted by participants. The supplied low, moderate, and high labels were used without relabeling; because the retained dataset did not include the original DASS-21 item level scoring record, this study reproduces prediction from the supplied categories rather than independently revalidating their psychometric scoring. Median imputation, scaling, and SMOTE were fitted only on each training partition, and test fold observations were excluded from all preprocessing estimates to prevent information leakage.

Table 1. Variables used in the prediction model

No.	Variable	Type	Meaning
1	Total screen time	Numerical	Total daily device use duration
2	Social media duration	Numerical	Daily social media use
3	Academic duration	Numerical	Device use for academic work
4	Entertainment duration	Numerical	Device use for entertainment
5	Notification frequency	Ordinal	Intensity of checking notifications
6	Late night use	Ordinal	Intensity of late night device use
7	Stress level	Target	DASS-21 derived low, moderate, or high

2.3. Stress labeling and class distribution

DASS-21 was used to establish the supervised learning target independently from the six digital activity predictors. Only the stress category was used as the target of the present model; the application does not claim to predict depression or anxiety. The aggregated cross validation confusion matrix also reveals the original class distribution: 43 students were labeled low stress, 15 moderate stress, and eight high stress. Consequently, 65.2% of the observations belonged to the low class, while only 12.1% belonged to the high class. Preserving these proportions in every fold was the reason for using stratification.

Table 2. Distribution of DASS-21 derived stress classes

No.	**Stress class**	**Number**	**Proportion**
1	Low	43	65.2%
2	Moderate	15	22.7%
3	High	8	12.1%
4	Total	66	100.0%

The distribution indicates a meaningful imbalance rather than an even three class problem. A majority classifier obtained 0.6516 ± 0.0347 accuracy by predicting every observation as low stress, but its macro F1-score was only 0.2629 ± 0.0084 and its recall for moderate and high stress was zero. Accordingly, accuracy and class balanced metrics are reported together.

2.4. Model development and evaluation

Random Forest was retained as the principal nonlinear model. Internal uncertainty was estimated with RepeatedStratifiedKFold using five splits, ten repeats, and random_state = 42, yielding 50 outer test folds. Five folds were chosen because the smallest class contained only eight observations; ten folds could create extremely small or empty class specific test subsets. All candidate algorithms used the same outer partitions. Scaling for Logistic Regression and SVM and SMOTE resampling for the selected imbalance aware Random Forest were confined to the training portion of each fold through a pipeline, preventing information leakage into the corresponding test fold.

The final imbalance aware pipeline was executed in Python 3.12.13 using scikit learn 1.6.1, imbalanced learn 0.14.2, pandas 2.2.2, and NumPy 2.0.2. The pipeline comprised median imputation, StandardScaler, SMOTE (sampling_strategy = 'auto', k_neighbors = 3, random_state = 42), and RandomForestClassifier (n_estimators = 100, criterion = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, max_features = 'sqrt', bootstrap = True, class_weight = None, random_state = 42, and n_jobs = None). The remaining estimator parameters retained the defaults documented by these library versions. No hyperparameter search was conducted; these settings constitute the final evaluated configuration

2.5. Evaluation measures

For each class, precision quantified the proportion of predictions assigned to that class that were correct, whereas recall quantified the proportion of actual members of the class that the model detected. F1-score was the harmonic mean of precision and recall. Macro averaging calculated each metric independently for low, moderate, and high stress and then gave the classes equal weight. This treatment was selected because it prevents the 43 member low class from completely dominating the evaluation (Yang & Berdine, 2024).

The confusion matrix was interpreted with rows representing actual labels and columns representing predicted labels. True predictions occupy the diagonal, while off diagonal values identify the direction of error. In an early screening context, confusion between moderate and low stress deserves particular attention because it may reduce the visibility of students who could benefit from additional support. Nevertheless, the study did not define clinical costs or use the model to initiate an intervention automatically. Repeated validation results are reported as mean ± standard deviation across the 50 test folds. These standard deviations describe variation across repeated partitions and are not treated as confidence intervals because folds overlap and are statistically dependent.

2.6. Application and testing

Python, Pandas, scikit learn, and Streamlit were used to implement StresCheck. The interface validates six inputs, generates a stress class, and displays level specific educational guidance and a disclaimer. Six black box scenarios tested navigation, invalid and valid input, prediction, reset, and session completion. Seven convenience users provided formative interface feedback; this was not treated as a validated usability study.

The available user evaluation record comprised ten questionnaire items (Q1–Q10) scored numerically on a five point scale by seven convenience users after they interacted with the prototype. Item level descriptive statistics and internal consistency were calculated. The retained record did not include the exact item wording or source, verbal response anchors, participant characteristics, scoring formula for the earlier percentage, or the raw 7 × 10 response matrix. Accordingly, the analysis is reported as preliminary formative feedback rather than a validated usability assessment.

2.7. System interaction and validation logic

The user first opens the home page and selects the prediction function. Four duration fields and two intensity fields are then completed. Before inference, the application checks whether the combined social

media, academic, and entertainment durations exceed total screen time. An inconsistent entry is rejected and accompanied by an error message; a valid entry is transformed into the feature order expected by the trained model. The predicted class controls the result page profile and the educational recommendations that follow.

After viewing the result, the user can restart the assessment or end the session. Restarting clears the current input state and returns to the form; ending the session opens a closing page with follow up guidance. The prototype does not store a longitudinal personal record and does not provide authenticated clinical access. These boundaries reduce functionality but also limit the privacy exposure of the present laboratory implementation.

3. RESULTS AND DISCUSSION

3.1. System implementation

StresCheck implemented a home page, digital activity form, logical validation, three result states, recommendation cards, repeat prediction, and session completion. Figure 2 presents the principal user interaction interfaces, whereas Figure 3 shows the result pages generated for the low, moderate, and high stress classes. The outputs are explicitly framed as educational screening information rather than clinical diagnoses.

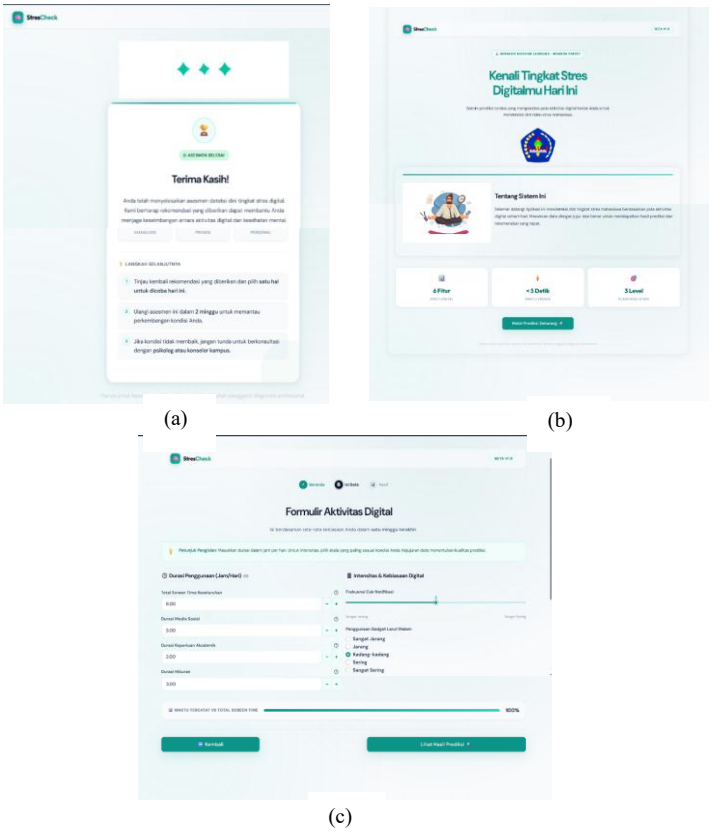
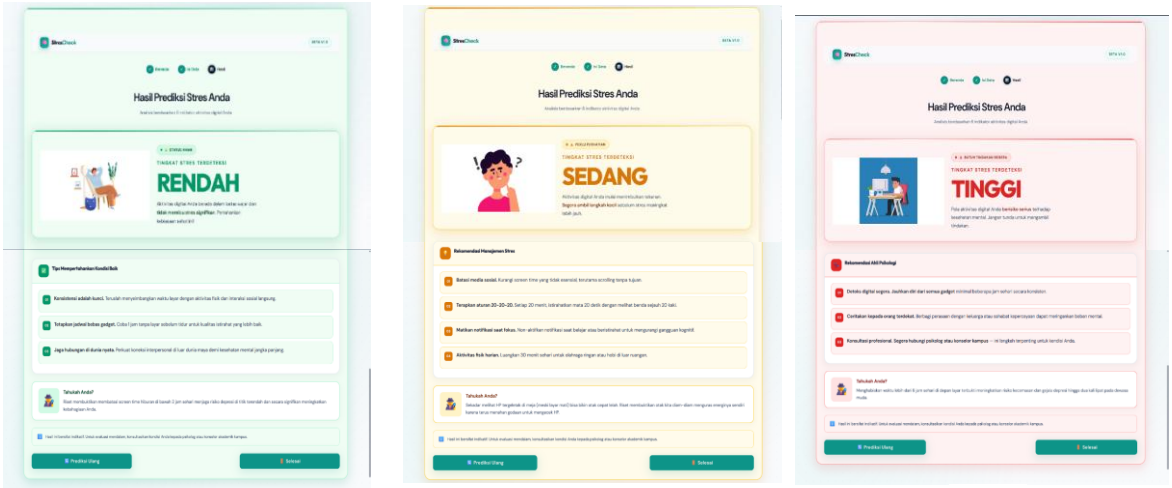


Figure 2. User interaction interfaces of StresCheck: (a) home page, (b) digital activity input form, and (c) session completion page.



(a) (b) (c)

Figure 3. Stress screening result interfaces generated by StresCheck: (a) low stress result, (b) moderate stress result, and (c) high stress result.

3.2. Classification performance

The metric audit regenerated the out of fold confusion matrix using the explicit label order Low, Moderate, High (Figure 4). Of 66 observations, 41 were correctly classified: 34 low, two moderate, and five high. Twelve of 15 actual moderate cases were assigned to low stress, one to high stress, and only two were correctly detected. This audited matrix replaces the earlier inconsistent values.

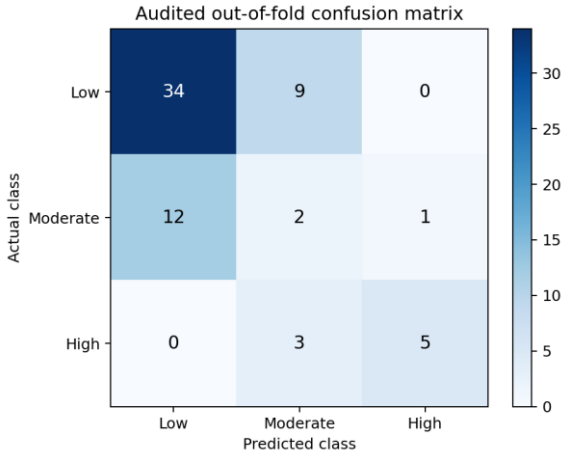


Figure 4. Audited out of fold confusion matrix (label order: Low, Moderate, High).

Table 3. Audited single out of fold Random Forest performance

No.	Metric	Value
1	Accuracy	62.12%
2	Macro precision	57.18%
3	Macro recall	51.63%
4	Macro F1-score	53.88%

The audited 62.12% accuracy is 3.04 percentage points below the repeated majority baseline mean of 65.16%; consequently, accuracy does not support a claim of superiority. The more relevant evidence is class balanced performance: the Random Forest out of fold macro F1-score was 53.88%, whereas the

repeated DummyClassifier macro F1-score was 26.29 ± 0.84% and its moderate and high class recall were both zero. This distinction prevents the dominant low stress class from creating an overly favorable interpretation.

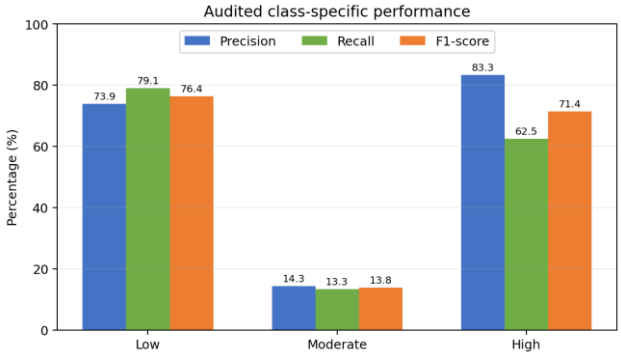


Figure 5. Audited precision, recall, and F1-score by stress class.

Table 4. Audited performance metrics for each stress class

Class	Precision	Recall	F1-score
Low	73.91%	79.07%	76.40%
Moderate	14.29%	13.33%	13.79%
High	83.33%	62.50%	71.43%

The high class achieved the highest precision (83.33%), but its 62.50% recall means that three of eight high stress observations were missed and assigned to moderate stress. The low class achieved 73.91% precision and 79.07% recall. Moderate stress performance was the principal failure: only two of 15 cases were correctly detected, producing precision of 14.29%, recall of 13.33%, and F1-score of 13.79%.

The diagnostic plots showed overlapping predictor distributions, particularly between low and moderate stress. For total screen time, the low group had a median of 7.0 hours (IQR 3.75), compared with 10.0 hours (IQR 4.24) for the moderate group. Thirteen of 15 moderate cases were described as close or ambiguous cases in the prediction audit; for example, case index 21 received equal probabilities of 0.47 for low and moderate stress. These observations indicate limited feature separability and uncertain decision boundaries. They do not prove label error, because raw DASS-21 scores near category cut offs were not supplied.

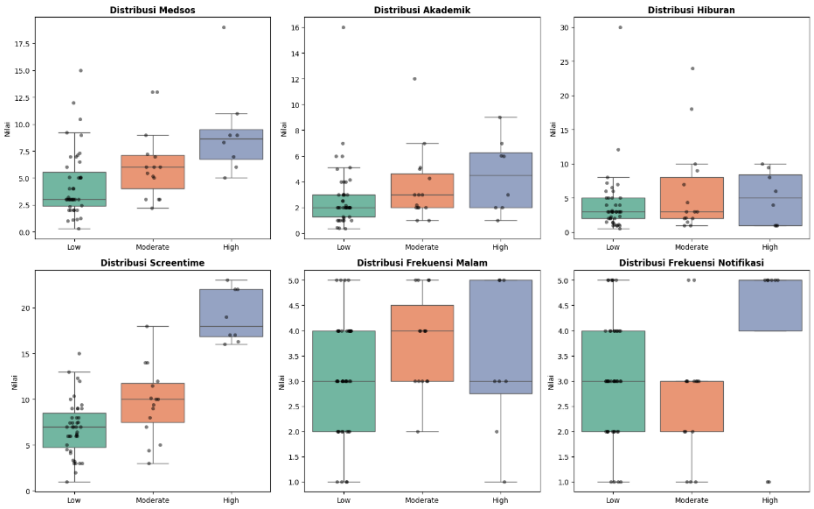


Figure 6. Predictor distributions across low, moderate, and high stress classes.

3.3. Repeated validation and comparative modeling

Across 50 repeated stratified test folds, the selected SMOTE Random Forest pipeline achieved accuracy of 0.6197 ± 0.1092, macro precision of 0.5769 ± 0.1115, macro recall of 0.5801 ± 0.1127, and macro F1-score of 0.5582 ± 0.0987. Class specific recall was 0.7336 ± 0.1659 for low, 0.2067 ± 0.2394 for moderate, and 0.8000 ± 0.3000 for high stress. The large standard deviations, especially for minority class recall, show that performance was sensitive to partition composition.

Table 5. Repeated stratified five-fold performance of the selected SMOTE–Random Forest

Metric	Mean	SD
Accuracy	0.6197	0.1092
Macro precision	0.5769	0.1115
Macro recall	0.5801	0.1127
Macro F1-score	0.5582	0.0987
Recall—Low	0.7336	0.1659
Recall—Moderate	0.2067	0.2394
Recall—High	0.8000	0.3000

Logistic Regression had the highest accuracy among the non dummy comparison models (0.7089) and macro F1 = 0.5523. The imbalance aware SMOTE Random Forest was retained because the study prioritized macro F1 and minority class recall; its macro F1 = 0.5582 was marginally higher. This narrow difference does not establish decisive algorithmic superiority.

Table 6. Comparison of baseline algorithms under the common validation design

Model	Accuracy	Macro F1	Recall Moderate	Recall High
Dummy	0.6516	0.2629	0.0000	0.0000
Logistic Regression	0.7089	0.5523	0.1933	0.6400
Decision Tree	0.6153	0.5406	0.2467	0.6700
SVM	0.6787	0.3773	0.0000	0.3100
Random Forest	0.6366	0.5259	0.1133	0.7300

Among the Random Forest imbalance treatments, SMOTE produced the highest macro F1 (0.5582) and moderate/high recall (0.2067/0.8000). Class weighting slightly improved macro F1 but reduced moderate recall to 0.08. The improvement should be interpreted cautiously because the moderate class SD remained large and precision recall trade-offs were not fully documented in the supplied output.

The observed trade off is consistent with comparative evidence that no rebalancing method is uniformly superior and that sensitivity gains may be accompanied by lower overall accuracy or unstable synthetic samples (Abdulsadig & Rodriguez-Villegas, 2024). Consequently, the present selection of

SMOTE–Random Forest is based on macro F1-score and minority-class recall within this dataset, not on a claim that SMOTE is generally optimal.

Table 7. Random Forest class-imbalance experiments

Configuration	Macro F1	Recall Moderate	Recall High
RF baseline	0.5259	0.1133	0.7300
Class weight: balanced	0.5417	0.0800	0.7500
Class weight: balanced_subsample	0.5453	0.0800	0.7700
SMOTE–RF	0.5582	0.2067	0.8000

3.4. Feature importance

Total screen time ranked first in both impurity based importance (0.503) and validation based permutation importance. Permuting total screen time decreased macro F1 by 0.2389 ± 0.1412. The other five features had means near or below zero, suggesting that they added little stable predictive information beyond the remaining variables in the validation folds. Negative permutation values do not prove that a feature is intrinsically harmful; they can arise from sampling variation, redundancy, or unstable small-sample estimates.

Agreement between impurity and permutation rankings was therefore partial: both placed total screen time first, but impurity importance assigned positive shares to every predictor whereas permutation importance did not confirm stable out of sample contributions for the remaining features. SHAP was not added because no validated SHAP output was supplied.

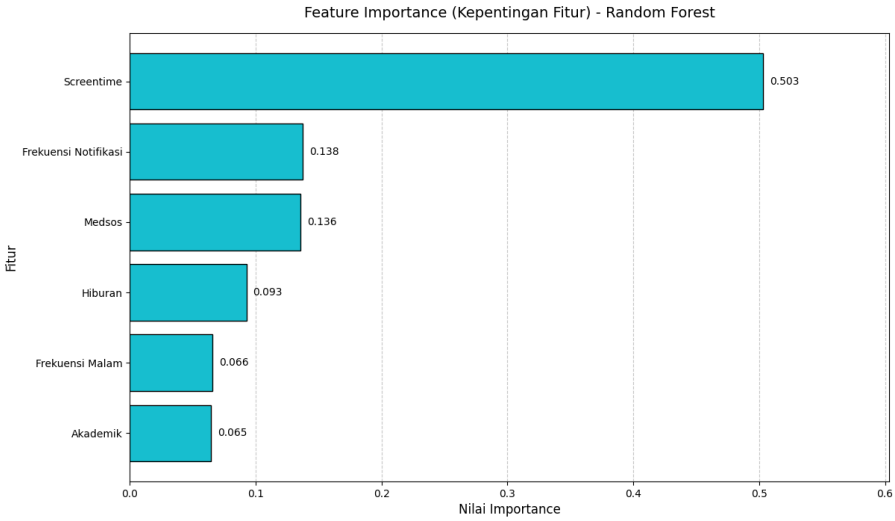


Figure 7. Impurity-based Random Forest feature importance ranking

Table 8. Impurity-based Random Forest feature-importance values

Rank	Feature	Importance
1	Total screen time	0.503
2	Notification frequency	0.138
3	Social-media duration	0.136

4	Entertainment duration	0.088
5	Late-night use	0.071
6	Academic duration	0.063

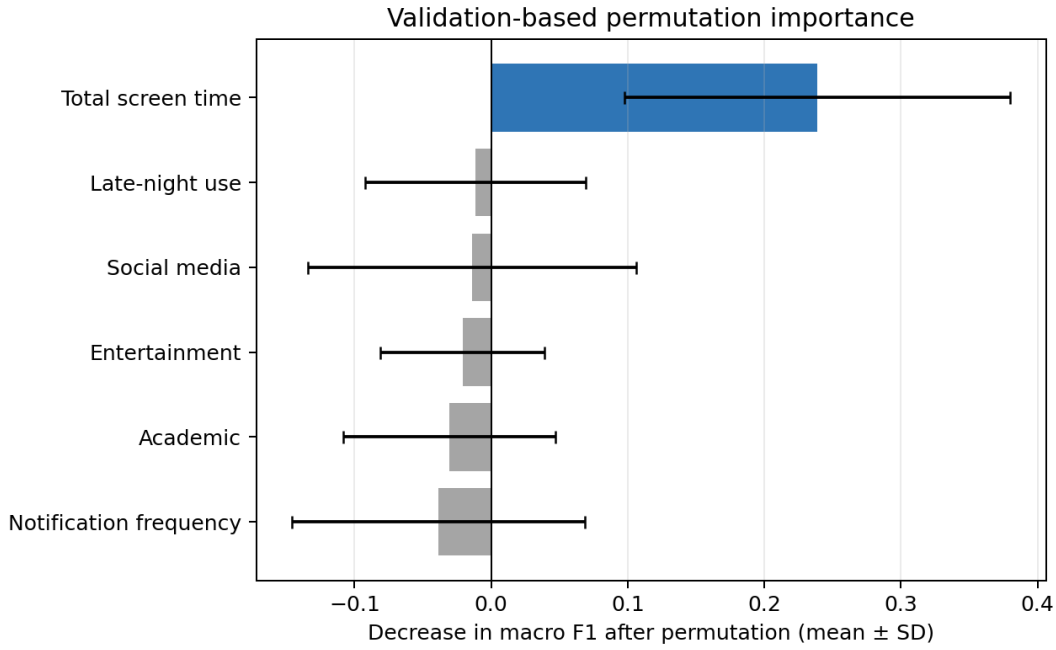


Figure 8. Validation-based permutation importance using macro F1.

Table 9. Validation-based permutation importance

Feature	Mean decrease in macro F1	SD
Total screen time	0.238870	0.141194
Late-night use	−0.011706	0.080695
Social-media duration	−0.014125	0.120098
Entertainment duration	−0.020780	0.060081
Academic duration	−0.030808	0.077638
Notification frequency	−0.038561	0.107051

3.5. Functional testing and preliminary user feedback

All six black box testing scenarios produced the expected results, demonstrating that the principal application functions operated according to the specified requirements. The scenarios covered the complete user journey and the main invalid input branch, as summarized in Table 10

Table 10. Black-box functional testing

Scenario	Expected behavior	Result
Open home page	Information and start button are displayed	Valid
Start prediction	System opens the six variable form	Valid
Invalid duration	Input is rejected and an error is displayed	Valid

Valid data	Prediction and recommendations are displayed	Valid
Repeat prediction	Inputs reset and form reopens	Valid
Finish session	Closing page and guidance are displayed	Valid

Across the ten recorded items, mean scores ranged from 3.86 (Q8) to 4.57 (Q3), with standard deviations from 0.53 to 1.13. Cronbach's alpha was 0.9186. Although this value indicates high internal consistency within the available responses, it is highly unstable with only seven users and cannot establish instrument validity. Table 11 reports the complete item level descriptive statistics available in the evaluation record; the earlier 86.3% satisfaction score is not used because its scoring formula could not be verified.

Table 11. Descriptive statistics for the ten-item user evaluation (n = 7)

Item	Mean	SD	Minimum	Maximum
Q1	4.29	0.95	3	5
Q2	4.43	1.13	2	5
Q3	4.57	0.79	3	5
Q4	4.43	0.79	3	5
Q5	4.43	0.53	4	5
Q6	4.29	0.95	3	5
Q7	4.14	0.90	3	5
Q8	3.86	0.90	3	5
Q9	4.29	0.76	3	5
Q10	4.43	0.98	3	5

3.6. Comparison with related studies

Table 12 provides a structured comparison with selected related studies. Sharif et al. (2023) used wearable physiological signals from 45 adults and reported 91% Random Forest accuracy, whereas Ku and Min (2024) examined the stability of five algorithms using a much larger student dataset under simulated subjective-response errors. Pereira et al. (2025) evaluated 274 psychology students using richer psychological and lifestyle variables and reported high performance for binary burnout-risk profiles. These results are not directly comparable with the present three-class task because the outcomes, predictors, sample sizes, class structures, and validation strategies differ.

Table 12. Comparison study

Study (sample)	Predictors / outcome	Algorithms	Validation / robustness	Reported performance	Comparability limitation
Sharif et al. (2023) 45 adults; 225 person-days	EEG, blood pressure, PANAS and personal variables; binary commute-related physiological stress label	RF, SVM, Naive Bayes, KNN	80:20 split; results with and without cross-validation were compared	RF accuracy 91%; SVM 80%; KNN 80%; NB 73%	Objective biosignals, repeated person-days, binary outcome, and different validation design
Ku & Min (2024) 4,184 students	Biomedical, demographic and self-reported survey variables; MDD and GAD prediction	CNN, RF, XGBoost, Logistic Regression, Naive Bayes	Ablation / simulated subjective-response error scenarios	Focus on model stability under increasing response error; no directly comparable three-class stress score	Different outcomes, substantially larger public dataset, and robustness-to-noise objective
Pereira et al. (2025) 274 students	Psychological and lifestyle variables; binary burnout-risk profiles	RF, XGBoost, SVM	Cross-sectional internal model evaluation; detailed split not used for direct comparison here	RF 95.06%; XGBoost 93.82%; SVM 97.53% accuracy	Binary burnout profiles, richer predictors, larger sample, and different outcome construction
Present study 66 students	Six self-reported digital-activity variables; three DASS-21-derived stress classes	Dummy, Logistic Regression, Decision Tree, SVM, Random Forest; SMOTE–RF selected	Repeated stratified five-fold CV (10 repeats); imbalance and sensitivity analyses	SMOTE–RF macro F1 0.5582 ± 0.0987; macro recall 0.5801 ± 0.1127	Small, imbalanced, single-program cohort; no external validation

The present findings also align with recent stress-modeling evidence in several important ways. Shvetcov et al. (2024) reported that passive smartphone sensing could distinguish severe stress but had difficulty characterizing mild to moderate cases, paralleling the weak moderate class separation observed here. Ratul et al. (2023) used a much larger survey of 444 university students and broader contextual variables, while Vos et al. (2023) showed that combining heterogeneous datasets improved generalization beyond models trained on small single study samples. The current 66 participant, six variable design therefore trades breadth and external validity for low respondent burden. Systematic reviews further show that stress applications vary greatly in data type and preprocessing and that rigorous reporting is necessary for reproducibility (Daza et al., 2023; Mittal et al., 2022). An open access scoping review likewise identifies interpretability, personalization, naturalistic settings, and real-time processing as priorities for stress-related machine learning (Razavi et al., 2024), whereas StresCheck remains limited to non diagnostic exploratory screening.

The main value of StresCheck is therefore not superior predictive performance, but the integration of an interpretable baseline model into a simple student facing workflow. Its moderate class failure provides a concrete research target. Adding sleep quality, academic workload, physical activity, and validated psychosocial variables may improve separability, but every additional feature increases respondent burden and privacy risk. Future design must balance accuracy, accessibility, and data minimization.

3.7. Implication and limitations

The results show that compact self-reported digital behavior can support experimental model development, but predictive performance depended strongly on validation design, feature composition, and class balance. The repeated estimates and learning curve indicate substantial uncertainty, while the algorithm comparison does not establish a decisive advantage for Random Forest.

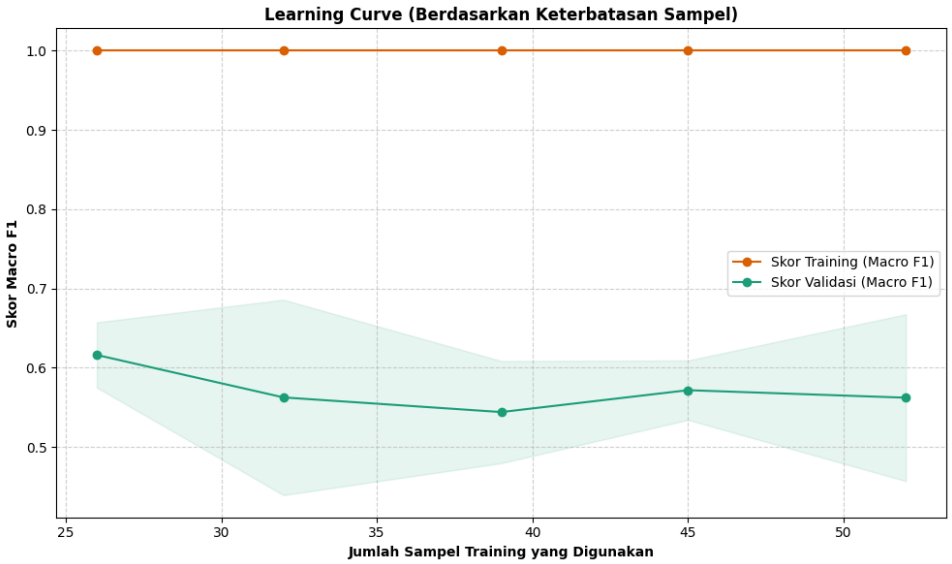


Figure 9. Learning curve showing the persistent training–validation performance gap.

Limitations include 66 participants from one program, severe class imbalance, self-reported activity, a DASS-21-derived categorical label, weak moderate-stress discrimination, and no external validation. The learning curve retained a perfect training macro F1 of 1.00 while validation remained approximately 0.54-0.62 and did not converge toward the training score. This persistent gap is consistent with high variance

and supports the need for a substantially larger, more diverse cohort; it does not determine a precise future sample requirement.

The seven-user evaluation is additionally limited by convenience sampling, the very small sample, and the absence of retained item wording, response anchor labels, participant characteristics, and the raw response matrix. The reported Cronbach's alpha and item summaries should therefore be interpreted only as preliminary descriptive evidence. Future evaluation should use a documented and validated instrument, prespecified scoring, a larger representative sample, and complete anonymized item level records.

A further procedural limitation is that prospective formal ethical review and documented informed consent for research publication were not obtained. This omission cannot be corrected by claiming retrospective approval and limits the strength of the present evidence. Future studies will seek institutional ethical review or a written exemption determination before recruitment; use documented informed consent; separate recruitment and consent from lecturers responsible for assessment; state clearly that participation or refusal has no academic consequences; and specify de-identification, access control, data retention, and withdrawal procedures.

3.8. Deployment considerations

A campus implementation must separate educational screening from clinical or administrative decision-making. The dominant error moderate stress predicted as low could provide false reassurance and delay help seeking, while missed high-stress cases could have more serious consequences. Therefore, a low model prediction must not override a student's self reported distress. Any future interface should display persistent support and referral information, communicate uncertainty, avoid stigmatizing labels, prohibit automated adverse action, and require qualified human oversight. The present prototype is not ready for institution wide deployment.

Data governance is equally important because digital routines and mental health labels are sensitive. Future versions should minimize collected identifiers, obtain explicit informed consent, encrypt stored records, restrict access by role, define retention periods, and allow participants to withdraw their data. Model monitoring should test whether errors differ across demographic groups. These requirements are consistent with the fairness, privacy, transparency, and interpretability priorities identified in recent student mental health machine learning reviews (Drira et al., 2024).

From a modeling perspective, the moderate class deserves targeted improvement. Additional samples may be collected through class aware recruitment, while class weighting, balanced resampling, ordinal classification, and probability calibration can be compared under nested cross validation. Any improvement should be evaluated using macro F1-score, per class recall, calibration error, and an untouched external cohort rather than accuracy alone.

4. CONCLUSIONS AND RECOMMENDATIONS

This exploratory study developed a technically functional prototype for predicting three DASS-21-derived stress categories from six self reported digital-activity variables. After metric auditing, the single out of fold evaluation yielded 62.12% accuracy and 53.88% macro F1. Under repeated stratified five-fold validation, the selected SMOTE–Random Forest achieved 61.97 ± 10.92% accuracy, 58.01 ± 11.27% macro recall, and 55.82 ± 9.87% macro F1. The majority DummyClassifier achieved higher accuracy

(65.16 ± 3.47%) but only 26.29 ± 0.84% macro F1 and zero recall for moderate and high stress. Thus, any value of the model lies in improved class balanced detection rather than overall accuracy.

The principal scientific contribution is empirical rather than algorithmic. The study demonstrates that a low burden six variable design can improve class balanced detection over a majority rule but remains insufficient for stable three class screening, particularly for moderate stress. By reporting audited metrics, repeated-fold variation, alternative classifiers, imbalance treatments, predictor overlap, and permutation importance together, the study provides a reproducible boundary condition for future data minimal student-stress models.

Moderate stress recall remained low (20.67 ± 23.94% under repeated validation), predictor distributions overlapped, minority class estimates were unstable, and no external validation was performed. Logistic Regression and SMOTE Random Forest produced similar macro F1 values, so Random Forest superiority is not established. StresCheck should therefore be regarded as an exploratory proof of concept, not a diagnostic or deployment-ready institutional screening tool. Larger and more balanced multi institutional cohorts, independent external validation, calibrated probabilities, prospective usability evaluation, and verified ethical governance are required before campus implementation.

Future work should recruit a larger, multi-institutional and class aware cohort; retain item level DASS-21 and usability instrument records; obtain prospective ethical review and documented consent, compare calibrated ordinal and cost sensitive models under nested validation, and evaluate an untouched external cohort. Deployment research should additionally measure calibration, false negative consequences, usability, fairness across demographic groups, and the effectiveness of referral guidance before any institutional screening use.

REFERENCES

Abdulsadig, R. S., & Rodriguez-Villegas, E. (2024). A comparative study in class imbalance mitigation when working with physiological signals. *Frontiers in Digital Health*, *6*. https://doi.org/10.3389/fdgth.2024.1377165

Beny, B. (2026). Performance Analysis of Ensemble Learning Models in Heart Failure Prediction: Random Forest, AdaBoost, and XGBoost. *Telematika*, *19*(1), 1–10. https://doi.org/10.35671/telematika.v19i1.3218

Boers, E., Afzali, M. H., Newton, N., & Conrod, P. (2019). Association of Screen Time and Depression in Adolescence. In *JAMA Pediatrics* (Vol. 173, Number 9, pp. 853–859). American Medical Association. https://doi.org/10.1001/jamapediatrics.2019.1759

Currey, D., & Torous, J. (2022). Digital Phenotyping Data to Predict Symptom Improvement and App Personalization: Protocol for a Prospective Study. *JMIR Research Protocols*, *11*(11). https://doi.org/10.2196/37954

Daud, N. M. (2025). From innovation to stress: analyzing hybrid technology adoption and its role in technostress among students. *International Journal of Educational Technology in Higher Education*, *22*(1). https://doi.org/10.1186/s41239-025-00529-x

Daza, A., Saboya, N., Necochea-Chamorro, J. I., Zavaleta Ramos, K., & Vásquez Valencia, Y. del R. (2023). Systematic review of machine learning techniques to predict anxiety and stress in college students. In *Informatics in Medicine Unlocked* (Vol. 43). Elsevier Ltd. https://doi.org/10.1016/j.imu.2023.101391

Defit, S., Windarto, A. P., & Alkhairi, P. (2024). Comparative Analysis of Classification Methods in Sentiment Analysis: The Impact of Feature Selection and Ensemble Techniques Optimization. *Telematika*, *17*(1), 52–67. https://doi.org/10.35671/telematika.v17i1.2824

Dessauvagie, A. S., Dang, H. M., Nguyen, T. A. T., & Groen, G. (2022). Mental Health of University Students in Southeastern Asia: A Systematic Review. In *Asia-Pacific Journal of Public Health* (Vol. 34, Numbers 2–3, pp. 172–181). SAGE Publications Inc. https://doi.org/10.1177/10105395211055545

Drira, M., Ben Hassine, S., Zhang, M., & Smith, S. (2024). Machine Learning Methods in Student Mental Health Research: An Ethics-Centered Systematic Literature Review. In *Applied Sciences (Switzerland)* (Vol. 14, Number 24). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/app142411738

Hilda, & Roy Purwanto, M. (2024). PENGARUH MEDIA SOSIAL TERHADAP KESEJAHTERAAN MENTAL MAHASISWA: STUDI KASUS DI FAKULTAS ILMU AGAMA ISLAMA UNIVERSITAS ISLAM INDONESIA. *At-Thullab : Jurnal Mahasiswa Studi Islam*, *6*(1), 1485–1486. https://doi.org/10.20885/tullab.vol6.iss1.art2

Jamilah, F. N. (2026). IMPLEMENTASI DATA MINING UNTUK PREDIKSI DAN KLASIFIKASI TINGKAT STRES MENGGUNAKAN ALGORITMA RANDOM FOREST. *Jurnal Informatika Dan Teknik Elektro Terapan*, *14*(1). https://doi.org/10.23960/jitet.v14i1.8838

Ku, W. L., & Min, H. (2024). Evaluating Machine Learning Stability in Predicting Depression and Anxiety Amidst Subjective Response Errors. *Healthcare (Switzerland)*, *12*(6). https://doi.org/10.3390/healthcare12060625

Malik, M., & Javed, S. (2021). Perceived stress among university students in Oman during COVID-19-induced e-learning. *Middle East Current Psychiatry*, *28*(1). https://doi.org/10.1186/s43045-021-00131-7

Mittal, S., Mahendra, S., Sanap, V., & Churi, P. (2022). How can machine learning be used in stress management: A systematic literature review of applications in workplaces and education. In *International Journal of Information Management Data Insights* (Vol. 2, Number 2). Elsevier B.V. https://doi.org/10.1016/j.jjimei.2022.100110

Pereira, M. G., Santos, M., Magalhães, R., Rodrigues, C., Araújo, O., & Durães, D. (2025). Burnout Risk Profiles in Psychology Students: An Exploratory Study with Machine Learning. *Behavioral Sciences*, *15*(4). https://doi.org/10.3390/bs15040505

Ratul, I. J., Nishat, M. M., Faisal, F., Sultana, S., Ahmed, A., & Al Mamun, M. A. (2023). Analyzing Perceived Psychological and Social Stress of University Students: A Machine Learning Approach. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e17307

Razavi, M., Ziyadidegan, S., Mahmoudzadeh, A., Kazeminasab, S., Baharlouei, E., Janfaza, V., Jahromi, R., & Sasangohar, F. (2024). Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review. In *JMIR Mental Health* (Vol. 11). JMIR Publications Inc. https://doi.org/10.2196/53714

Sa'diyah, M., Naskiyah, N., & Rosyadi, A. R. (2022). Hubungan Intensitas Penggunaan Media Sosial Dengan Kesehatan Mental Mahasiswa Dalam Pendidikan Agama Islam. *Edukasi Islami: Jurnal Pendidikan Islam*, *11*(03), 713. https://doi.org/10.30868/ei.v11i03.2802

Sharif, M. S., Raj Theeng Tamang, M., Fu, C. H. Y., Baker, A., Alzahrani, A. I., & Alalwan, N. (2023). An Innovative Random-Forest-Based Model to Assess the Health Impacts of Regular Commuting Using Non-Invasive Wearable Sensors. *Sensors*, *23*(6). https://doi.org/10.3390/s23063274

Shvetcov, A., Funke Kupper, J., Zheng, W. Y., Slade, A., Han, J., Whitton, A., Spoelma, M., Hoon, L., Mouzakis, K., Vasa, R., Gupta, S., Venkatesh, S., Newby, J., & Christensen, H. (2024). Passive sensing data predicts stress in university students: a supervised machine learning method for digital phenotyping. *Frontiers in Psychiatry*, *15*. https://doi.org/10.3389/fpsyt.2024.1422027

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. In *Electronics (Switzerland)* (Vol. 12, Number 8). MDPI. https://doi.org/10.3390/electronics12081789

Twenge, J. M., & Campbell, W. K. (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive Medicine Reports*, *12*, 271–283. https://doi.org/10.1016/j.pmedr.2018.10.003

Vos, G., Trinh, K., Sarnyai, Z., & Rahimi Azghadi, M. (2023). Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical Informatics*, *148*. https://doi.org/10.1016/j.jbi.2023.104556

Yang, S., & Berdine, G. (2024). Confusion matrix. *The Southwest Respiratory and Critical Care Chronicles*, *12*(53), 75–79. https://doi.org/10.12746/swrccc.v12i53.1391