International Journal of Advances in Data and Information Systems
Vol. 7, No. 3, December 2026, pp. 1011~1029
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i3.1642

Cost-Sensitive DASS-42 Risk Classification in Young Adults Using Asymmetric Misclassification Costs

Iklima Fatma Aliena[1], Riska Dhenabayu[2], Rully Mujiastuti[3], Tatiana Starostina[4], Ratih Amelia[5]
[1,2,5]Department of Digital Business, Universitas Negeri Surabaya, Indonesia
[3]Department of Informatics Engineering, Universitas Muhammadiyah Jakarta, Indonesia
[4]Department of Texts, Languages and Digital Tools, Higher School of Economics University, Russia

Article Info

Article history:

Received Jun 13, 2026
Revised Aug 21, 2026
Accepted Sept 18, 2026

Keywords:

Asymmetric Misclassification
DASS-42
Cost-Sensitive Learning
Mental Health Classification
Young Adults

ABSTRACT

Young adults experience elevated mental health risk, yet most DASS-42 machine learning studies use heterogeneous populations, single-split validation, and accuracy-driven evaluation that treats clinically harmful errors as equal. This study developed a cost-sensitive machine learning framework for three-class mental health risk screening among young adults aged 18–25. A total of 23,039 DASS-42 records were filtered from an international survey and labeled as Low, Moderate, or High risk using the Maximum Severity Rule across depression, anxiety, and stress scores. Six models Naïve Bayes, Decision Tree, Random Forest, XGBoost, LightGBM, and Artificial Neural Network were compared. Class imbalance was handled using class weighting, while Cost-Proportionate Weighting imposed larger penalties on high-risk underclassification through an asymmetric cost matrix. Six models were evaluated using 10-fold stratified cross-validation and held-out testing, with nested hyperparameter tuning for the four tree-based models and fold-wise early stopping for the neural network. Performance was measured using Macro F1-score, ROC-AUC, confusion matrices, and total misclassification cost. The neural network achieved the strongest baseline performance (Macro F1 = 0.9686). Under CPW, its total CSL cost fell to 131, the lowest among all six models. CPW cut cost across all models, from 5.4% to 31.05%, with the neural network showing the largest reduction. Five of six models produced zero Low-High errors; only Decision Tree missed that mark. A rule-based scoring classifier reached 100% accuracy by construction, indicating near-perfect model scores reflect reconstruction of the scoring rule rather than genuine clinical prediction. These findings support cost-aware DASS-42-based screening for early risk identification.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Riska Dhenabayu,
Department of Digital Business,
Universitas Negeri Surabaya,
Jl. Ketintang, Gayungan, Surabaya, Jawa Timur, Indonesia, 60231.
Email: riskadhenabayu@unesa.ac.id

1. INTRODUCTION

Mental health disorders have become one of the most pressing public health issues of the twenty-first century. In 2019, the World Health Organization estimated that there were approximately 970 million people worldwide with a mental disorder, among them 301 million with anxiety disorders and 280 million with depression [1]. These conditions impose substantial individual and societal burdens, reducing productivity, impairing daily functioning, and significantly

diminishing quality of life [2]. Young adults aged 18 to 25 are disproportionately affected. This age group often faces overlapping academic, employment, and independent-living transitions [3]. Systematic review and meta analysis of 64 studies found that depression and anxiety symptoms affect roughly one in three college students, with pooled prevalence rates of 33.6% and 39.0% respectively [4], [5]. Yet despite bearing this disproportionate burden, young adults remain among the least served by existing mental health systems. A separate meta-analysis found that college students with mental health conditions consistently avoid seeking help, citing stigma, misconceptions about treatment, and structural access barriers as the most common reasons [6].

Machine learning has opened up new approaches to automated risk classification. Researchers have applied supervised learning algorithms to predict psychological disorders from structured self-report data, and results across different instruments and populations have been encouraging [7], [8], [9]. The Depression Anxiety Stress Scales-42 (DASS-42) has become a widely adopted data source for this purpose, offering standardized and multidimensional measurements across three clinically distinct domains: depression, anxiety, and stress [10]. Models including Random Forest, Support Vector Machine, and Artificial Neural Network have achieved strong classification performance on DASS-42 data [11], [12], [13]. Biswas et al. [14] found that a neural network applied to the full DASS-42 dataset reached 99.73 percent accuracy for depression classification. Reviews of AI applications in mental health confirm that deep learning models have demonstrated strong performance on psychological data [15], [16].

Models trained on heterogeneous age groups may miss response patterns specific to high-risk subgroups like young adults, whose psychological profiles differ from those of older adults [3]. Biswas et al. [14] acknowledged this and recommended that future work target specific at-risk demographics. Several prior studies have relied on limited validation protocols rather than rigorous cross-validation, which may overestimate model performance [8]. Biswas et al. [14] recommended k-fold cross-validation to verify model stability. A third problem: conventional metrics like accuracy and F1-score treat all misclassification errors as equally costly. In clinical screening, missing a high-risk individual carries far greater consequences than the reverse [17], [18]. The absence of cost-sensitive training and evaluation in most DASS-42 studies limits how applicable those models are in practice.

Two questions drive this study. Which machine learning model produces the highest and most stable classification performance for mental health risk prediction in adults aged 18 to 25, using 10-fold stratified cross-validation on the DASS-42 instrument? And does Cost-Proportionate Weighting reduce total misclassification cost compared to standard class balancing, and if so, which models benefit most? To answer both questions, this study compares six models, namely Naive Bayes, Decision Tree, Random Forest, XGBoost, LightGBM, and Artificial Neural Network, on a three-class mental health risk classification task covering 23,039 young adult respondents from the DASS-42 dataset. Three methodological choices set this work apart from prior studies. The dataset was filtered to the 18–25 age group, addressing the demographic gap Biswas et al. [14] flagged. All six models were evaluated using 10-fold stratified cross-validation on the training set. Hyperparameters for Decision Tree, Random Forest, XGBoost, and LightGBM were tuned through a nested inner-loop search to avoid optimistic bias. The ANN, whose architecture was fixed by design, followed the same 10-fold cross-validation protocol, with early stopping monitored independently within each fold. Cost-Proportionate Weighting (CPW) [19] was integrated into the training pipeline to penalize the most clinically harmful errors. Total misclassification cost, derived from an asymmetric cost matrix, serves as a complementary metric alongside Macro F1-score and area under the ROC curve. Together, these choices constitute the primary novelty of this work relative to existing DASS-42 studies [11], [12], [14]. It is important to note that the proposed framework is intended for DASS-42-based screening support rather than clinical diagnosis. Because the risk labels are derived from DASS-42 item responses, the contribution lies in cost-sensitive, demographically targeted, and clinically weighted risk classification rather than prediction of externally confirmed psychiatric diagnoses.

A further issue in this literature, one largely unexamined so far, concerns how labels are constructed. In DASS-42 studies including Biswas et al. [14], Kumar et al. [11], and Yang et al. [12], the three-class or five-class risk label comes directly from the same item responses used as model

features, computed through a deterministic scoring rule [20]. This is a feature-to-target leakage pattern documented in the broader machine learning literature [21], where near-perfect model performance reflects reconstruction of a known deterministic function rather than genuine generalization to an independently observed outcome [22]. No prior DASS-42 study, to our knowledge, has tested or reported this circularity. This study addresses the gap directly. Alongside the six supervised models, we implement a rule-based deterministic classifier that applies the DASS-42 scoring manual without any learning, making the circularity transparent and quantifiable, and we reframe the study's contribution accordingly. Because the deterministic rule is error-free by construction on this exact task, no learned model can offer a practical advantage over it in either accuracy or cost for label reconstruction. Instead, we identify and empirically test a capability the rule structurally lacks: continuous, cross-domain-informed confidence estimates usable for triage of borderline cases (Section 3.8.6).

2. RESEARCH METHOD

This section describes the experimental protocol used to evaluate cost-sensitive DASS-42-based risk classification among young adults. The workflow consisted of six stages: dataset screening, DASS-42 item preprocessing, three-class global risk label construction, model training and hyperparameter tuning, class-imbalance and cost-sensitive weighting, and final evaluation using both conventional and cost-sensitive metrics [19], [23]. Six supervised learning models were compared: Naive Bayes, Decision Tree, Random Forest, XGBoost, LightGBM, and Artificial Neural Network. All six models were evaluated using 10-fold Stratified K-Fold Cross-Validation on the training set. For Decision Tree, Random Forest, XGBoost, and LightGBM, hyperparameter tuning was nested within an inner cross-validation loop to avoid optimistic bias, while the ANN, whose architecture was fixed by design, was retrained with EarlyStopping independently within each outer fold. The held-out test set was not used during preprocessing parameter estimation, hyperparameter tuning, model selection, or threshold-independent evaluation [19], [24]. The full workflow is shown in Figure 1.

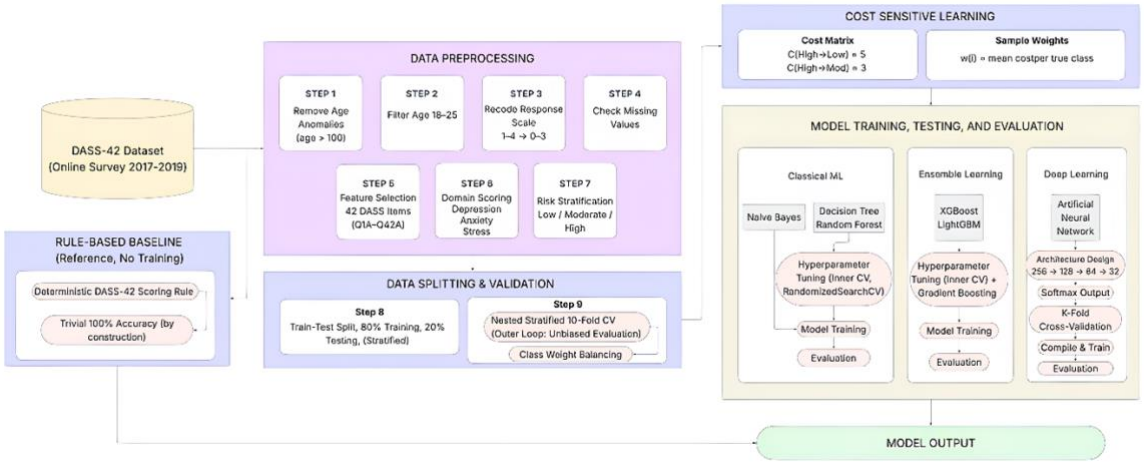


Figure 1. Proposed Methodology of Our Work

2.1. Data Source, Eligibility Criteria, and Ethical Consideration

The DASS-42, a self-report psychometric instrument created by Lovibond [20], was used in an international online survey conducted between 2017 and 2019 [25]. The raw dataset comprised 39,775 records. After removing records with anomalous age values and filtering to the 18–25 age range, 23,039 samples remained. Young adults in this age range carry a disproportionate burden of depression and anxiety, a pattern that recurs across multiple studies [3], [26], driven largely by academic pressure and financial instability during a period of identity formation, academic transition, and early career uncertainty [3]. Most prior DASS-based machine learning studies, including Biswas et al. [14], drew from the general population without age stratification, which limits how well their findings apply to this group. This study focuses specifically on the 18–25 subgroup to close that gap.

Because this study used a publicly available secondary dataset, only de-identified item responses and age information were analyzed. No personally identifiable information was used.

2.2. Data Preprocessing and Label Construction

Item responses were recoded from the original 1–4 response scale to the standard 0–3 DASS-42 scoring scale [20]. All 42 items had zero missing values, and only Q1A to Q42A were retained as model inputs. Domain scores for depression, anxiety, and stress were computed by summing the 14 corresponding items in each domain, consistent with the DASS-42 scoring manual [20]. Since the target label was derived directly from DASS-42 item scores, this study addresses automated risk classification based on self-report data rather than externally confirmed psychiatric diagnoses.

Each domain score was mapped to one of five severity levels: Normal, Mild, Moderate, Severe, and Extremely Severe. For the three-class global risk label, Normal and Mild were grouped as Low risk, Moderate remained Moderate risk, and Severe and Extremely Severe were grouped as High risk. The final label was assigned using the Maximum Severity Rule, where the highest severity level across the three domains determined the respondent's overall risk class.

Before training, StandardScaler normalization was fitted on the training set only to prevent data leakage [21]. During cross-validation, scaling was implemented within each fold so that the scaler was fitted only on the fold-specific training partition and then applied to the corresponding validation partition. The resulting distribution was: Low (n = 4,274; 18.55%), Moderate (n = 3,776; 16.39%), and High (n = 14,989; 65.06%).

2.3. Experimental Design, Train-Test Split, and Hyperparameter Optimization

Six supervised learning models were selected, ranging from simple probabilistic classifiers to deep learning. Naive Bayes is based on Bayes theorem with the assumption of independence between predictors [27]. Decision Tree partitions the feature space recursively using Gini or entropy criteria [28]. Random Forest aggregates individual predictions from decision trees via random resampling methods [29]. XGBoost builds gradient-boosted trees using second-order Taylor loss approximation [30]. LightGBM extends this approach with Gradient based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [31]. The Artificial Neural Network (ANN) is a multi-layer perceptron with ReLU activations, Batch Normalization, and Dropout regularization [32], built on layers of [256, 128, 64, 32] neurons with a three-class softmax output, trained with the Adam optimizer (learning rate = 0.001), EarlyStopping, and ReduceLROnPlateau callbacks over up to 150 epochs. Naive Bayes was trained using default parameters. Its one tunable hyperparameter, var_smoothing, has little influence on classification performance compared with the structural hyperparameters, such as tree depth, ensemble size, and boosting rate, that govern the other classical models. The dataset was split into 80% training (n = 18,431) and 20% testing (n = 4,608) using stratified sampling. Model stability was assessed with 10-fold Stratified K-Fold Cross-Validation on the training set. Consistent with standard supervised-learning evaluation practice, the *risk_global* label was withheld from the test partition during prediction and reintroduced only for post-hoc scoring against model outputs, ensuring no direct label access occurred at inference time. Table 1 presents the hyperparameter search spaces.

Table 1. Hyperparameter Search Spaces for Each Model

Model	Method	Key Hyperparameter Search Space
Naive Bayes	No tuning - default parameters	Default var smoothing: (1e-9)
Decision Tree	Nested RandomizedSearchCV (15 iter; inner 3-fold for tuning, outer 10-fold for unbiased evaluation; separate 10-fold search for final model)	max_depth: {5,10,15, None}, min_samples_split: {2,5,10}, min_samples_leaf: {1,2,4}, criterion: {gini, entropy}, class_weight: {balanced}
Random Forest	Nested RandomizedSearchCV (15 iter; inner 3-fold for tuning, outer 10-fold for unbiased evaluation; separate 10-fold search for final model)	n_estimators: {100,200,300}, max_depth: {None,10,20}, min_samples_split: {2,5}, min_samples_leaf: {1,2}, max_features: {sqrt}, class_weight: {balanced}
XGBoost	Nested RandomizedSearchCV (15 iter; inner 3-fold for tuning, outer 10-fold for unbiased	n_estimators: {100–300}, max_depth: {5,7,9}, learning_rate: {0.05,0.1,0.2}, subsample: {0.6,0.8,1.0},

	evaluation; separate 10-fold search for final model)	colsample_bytree: {0.6,0.8,1.0}
LightGBM	Nested RandomizedSearchCV (15 iter; inner 3-fold for tuning, outer 10-fold for unbiased evaluation; separate 10-fold search for final model)	n_estimators: {100–300}, learning_rate: {0.05,0.1,0.2}, num_leaves: {31,50,70}, subsample: {0.6,0.8,1.0}, colsample_bytree: {0.6,0.8,1.0}
ANN	Manual (150 epochs, EarlyStopping); K-Fold Cross-Validation (10-fold, fixed architecture, no hyperparameter search)	Architecture: [256,128,64,32,3], Dropout: [0.3,0.3,0.2,0.2], lr=0.001, ReduceLROnPlateau

For Decision Tree, Random Forest, XGBoost, and LightGBM, hyperparameter tuning and performance evaluation were nested to avoid optimistic bias. An outer 10-fold stratified cross-validation loop estimated generalization performance, while an inner 3-fold stratified loop, applied independently within each outer training fold, selected hyperparameters through RandomizedSearchCV. A separate hyperparameter search over the full training set, also using 10-fold cross-validation, produced the final deployed model used for held-out test evaluation. The ANN followed a matching 10-fold cross-validation protocol, with the fixed architecture retrained in each fold and early stopping applied independently, ensuring a comparable evaluation across all six models. Minor run-to-run variation in ANN metrics, typically under 1% in Macro F1, is expected due to non-deterministic GPU-based floating-point operations in TensorFlow/Keras, even with fixed random seeds. Classical models, from Naive Bayes through LightGBM, are fully deterministic under scikit-learn's *random_state* mechanism and reproduce identical results across runs.

2.4. Class Imbalance Handling

The final dataset was substantially imbalanced, with the High-risk class representing the largest proportion of respondents. Class weight balancing was applied as the baseline strategy to account for this, upweighting minority classes during training without altering the original data distribution or introducing synthetic samples.

As reviewed by Hida et al. [24] and Araf et al. [19], algorithm-level approaches are increasingly preferred over data-level oversampling for medical classification because they preserve the original response distribution. Class weight balancing was used as the class-imbalance baseline for models supporting class or sample weighting using Equation:

$$w_{(j)} = \frac{N}{K \times N(j)} \quad (1)$$

where w(j) is the weight for class j, N is total training samples, K = 3 (number of classes), and N(j) is the sample count for class j. The class weight for class j was computed as the total number of training samples divided by the product of the number of classes and the number of samples in class j. Resulting weights: Low = 1.797, Moderate = 2.034, High = 0.512. Unlike SMOTE, this approach introduces no synthetic data into the training process [17], [19]. Two training configurations were therefore compared: (1) a class-balanced baseline, designed to address class-frequency imbalance, and (2) a cost-sensitive configuration, designed to penalize clinically harmful errors using CPW. This separation was used to isolate the effect of asymmetric misclassification costs from ordinary class imbalance correction.

2.5. Asymmetric Cost Matrix and Cost-Proportionate Weighting

Class weight balancing corrects class-frequency imbalance but does not distinguish between clinically minor and clinically harmful errors. In mental health screening, underclassifying a High-risk respondent is more consequential than overclassifying a Low-risk respondent [17], [18]. Therefore, CPW was used to incorporate asymmetric misclassification costs into the training process. CPW is a preprocessing-level technique that converts a misclassification cost matrix into per-instance training weights [23], formalized by Zadrozny et al. [23] and applied in high-stakes medical contexts including emergency readmissions [33], [34], seizure detection [35], depression severity prediction [17], and myocardial infarction complications [36]. The misclassification cost matrix C was defined based on clinical reasoning, as presented in Table 2.

Table 2. Misclassification Cost Matrix

Actual/Predicted	Low	Moderate	High
Low (0)	0	1	2
Moderate (1)	2	0	1
High (2)	5	3	0

The cost values reflect clinical judgment. Misclassifying a High-risk individual as Low-risk (C = 5) carries the highest cost: it leaves someone at serious mental health risk without any intervention [17], [18]. Misclassifying High as Moderate (C = 3) is a significant underestimation that may lead to inadequate treatment. Over-predicting risk, Low classified as High (C = 2), wastes resources but does not directly harm the patient. Boundary misclassifications involving Moderate are assigned costs of 1 or 2 [37]. This structure follows the well-established clinical ML principle that false negatives on high-severity classes cost more than false positives [19], [23].

More generally, these costs follow an ordinal-distance structure, in which errors between adjacent severity classes receive lower costs than errors spanning the full severity range. The asymmetric inflation on High-to-Low errors described above intentionally departs from strict ordinal symmetry. A High-risk respondent misclassified as Low, for example, incurs a cost of 5, while the same respondent misclassified as Moderate incurs a cost of 3, encoding graduated severity rather than a binary correct or incorrect penalty. These values reflect the authors' clinical reasoning, informed by the cost-sensitive mental health screening literature [17], [18], [19], rather than direct consultation with a clinical expert panel. Empirical or expert-panel calibration of this matrix remains important future work (Section 4). Although the exact numerical values were not independently validated by a clinical panel, their ordinal structure is anchored in the DASS-42's own clinically validated severity hierarchy [20], under which the instrument's five severity levels were empirically derived and confirmed across multiple populations [10]. The cost matrix therefore encodes an existing clinical ordering rather than an arbitrary one, even though the precise magnitude of each cost remains a heuristic choice pending expert-panel calibration.

The per-instance sample weight under CPW is computed according to Equation (2):

$$w_{(Xi)} = \frac{1}{|C|-1} \sum_{j \neq y} C(y_i, j) \quad (2)$$

Here, x(i) is the i-th training sample, y(i) its true label, |C| = 3, and C(y(i), j) the cost of predicting class j when the true class is y(i). Weights were normalized to a mean of 1 [19], yielding: Low = 0.480, Moderate = 0.480, High = 1.279. This formulation converts the full pairwise cost matrix into a single per-instance weight, by averaging, for each true class, the costs of all possible misclassifications from that class [19], [23]. We chose this simplification for compatibility with the *sample_weight* interface of standard classifiers, which accepts one scalar weight per training instance rather than a full matrix of class-pair costs. A consequence of this design is that CPW cannot penalize different types of misclassification differently once training is complete. The model is encouraged to avoid errors from a costly true class in aggregate, but the sample weight itself does not distinguish, for example, a High-to-Moderate error from a High-to-Low error during training. The resulting weight is then passed to each classifier's *sample_weight* parameter, or, for the ANN, aggregated to class-level weights for the Keras *class_weight* parameter, during model fitting, so that instances belonging to classes with costlier potential errors contribute proportionally more to the training loss.

For the ANN, class-level CPW weights were derived from cost matrix row averages and passed to the Keras class_weight parameter: Low = 0.643, Moderate = 0.643, High = 1.714, following the weighted cross-entropy formulation in Araf et al. [19].

The total misclassification cost for evaluation is defined in Equation (3):

$$TC = \sum_i C(y^{(i)}, \hat{y}^{(i)}) \quad (3)$$

where TC is the total cost, y(i) is the true label, and ŷ(i) is the predicted label for test sample i. A lower TC indicates fewer costly misclassifications and greater suitability for cost-aware screening support [19], [36].

2.6. Evaluation Metrics and Error-Cost Analysis

Model performance was measured using classification performance metrics and cost-sensitive error metrics. Because the study focuses on asymmetric screening errors, special attention was given to High-risk recall, High-to-Low misclassification counts, Low-to-High misclassification counts, and total misclassification cost. Macro F1-score was the primary metric, assigning equal weight to all three classes regardless of frequency [14], [24]. The per-class F1-score is computed via Equation (4):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

The Macro F1-score is the mean of per-class F1-scores for all three risk categories. ROC-AUC was computed from predicted class probabilities using the One-vs-Rest (OvR) strategy per class, not from hard class labels [38], measuring each model's discriminative ability across decision thresholds. Confusion matrix analysis examined per-class misclassification patterns, with attention to High-to-Low errors as the most clinically critical error type. Total Misclassification Cost (TC), defined in Equation (3), quantifies screening error consequences under the CPW framework [19], [23]. A lower TC indicates fewer clinically costly errors. In addition to total misclassification cost, average misclassification cost can be reported as TC divided by the number of test samples, allowing comparison across future datasets with different sample sizes.

As an additional evaluation reference, we implemented a rule-based deterministic classifier that applies the DASS-42 scoring manual directly. It sums domain items, applies published thresholds, and selects the maximum severity across domains [20], without any learning involved. Because the risk label is constructed from this exact procedure, the classifier reaches 100% accuracy by construction. This provides a transparent, quantifiable reference point against which the practical scope and limits of the learned models are examined in Section 3.8.6.

3. RESULTS AND DISCUSSION

3.1. Dataset and Label Distribution

The dataset was collected between 2017 and 2019 through an international online survey [25] and originally contained 39,775 entries. After removing records with anomalous age values and filtering to the 18 to 25 age range, 23,039 samples remained. The final analytic sample therefore represented young adult respondents only, consistent with the study objective of developing an age-specific DASS-42 risk classification framework. Each record consists of 42 DASS item features (Q1A to Q42A) and one global risk label derived from the scoring procedure in Section 2.

Labels followed the DASS-42 clinical scoring manual [20]. The highest severity code across Depression, Anxiety, and Stress was mapped to one of three global risk classes: Low (0), Moderate (1), and High (2). Table 3 shows the resulting class distribution and stratified train-test split.

Table 3. Dataset Distribution and Train/Test Split

Risk Level	Count	Percentage	Split (Train/Test)	Class Weight
Low (0)	4,274	18.55%	3,419 / 855	1.797
Moderate (1)	3,776	16.39%	3,021 / 755	2.034
High (2)	14,989	65.06%	11,991 / 2,998	0.512
Total	23,039	100%	18,431 / 4,608	-

The imbalance in the dataset is shown in Table 3. The High-risk class accounts for 65.06% of all samples, likely reflecting the self-selection typical of open online psychological surveys, where people in distress are more inclined to participate[3], [14]. Class weight balancing was applied to all models during training. Oversampling techniques such as SMOTE were not used, as they introduce synthetic samples into the training process; class weighting is more conservative and keeps the original data distribution intact [19], [33].

3.2. Overall Model Performance

Following the preprocessing pipeline, label construction, and hyperparameter configurations described in Section 2, each model was evaluated on the held-out test set (n = 4,608) using Macro

F1-score as the primary metric, which assigns equal weight to all classes regardless of frequency and is appropriate for imbalanced multi-class distributions [24], [39]. 10-fold Stratified K-Fold Cross-Validation was applied to all six machine learning models on the training set to assess generalization stability. Table 4 presents the full performance summary. This subsection addresses the first research question by identifying the model with the strongest held-out classification performance and the most stable cross-validation behavior among the classical models.

Table 4. Model Performance Summary

Model	Test macro F1	Test Accuracy	10-Fold CV Macro F1	Stability Note
Naive Bayes	0.8174	0.85	0.8289 ±0.0082	Stable but lower F1
Decision Tree	0.7739	0.84	0.7646±0.0139	Weakest
Random Forest	0.8755	0.92	0.8821±0.0063	Moderate
XGBoost	0.9141	0.94	0.9175±0.0053	**Most stable**
LightGBM	0.9149	0.94	0.9183±0.0089	Strong, slightly higher mean
ANN	0.9686	0.98	**0.9680±0.0065**	Highest held-out performance

The ANN achieved the highest Macro F1-score on the test set at 0.9686, with a test accuracy of 0.98, and the highest 10-fold cross-validation Macro F1 among all six models at 0.9680 ± 0.0065, as shown in Table 4. Among the classical models, XGBoost showed the most stable cross-validation performance, with a CV mean of 0.9175 ± 0.0053 and the lowest standard deviation across all six models, including the ANN. LightGBM achieved a marginally higher cross-validation mean of 0.9183 ± 0.0089, though with greater variability. XGBoost's held-out Macro F1 of 0.9141 was comparable to LightGBM's 0.9149. Random Forest reached a Macro F1 of 0.8755. Naive Bayes and Decision Tree, included as baselines, obtained 0.8174 and 0.7739 respectively, consistent with reports that simpler probabilistic and single-tree classifiers tend to underperform on imbalanced multi-class psychometric data [8], [40].

3.3. Per-Class Performance Analysis

To examine model behavior across individual risk levels, per-class precision, recall, and F1-scores were extracted from the classification reports on the test set. Table 5 presents the per-class performance for all six models.

Table 5. Per-Class Precision, Recall, and F1-Score for All Models

Model	Low P/R/F1	Moderate P/R/F1	High P/R/F1
Naive Bayes	0.92/0.82/0.87	0.53/0.90/0.67	0.99/0.85/0.92
Decision Tree	0.81/0.81/0.81	0.52/0.68/0.59	0.96/0.89/0.92
Random Forest	0.90/0.89/0.90	0.74/0.78/0.76	0.97/0.96/0.97
XGBoost	0.94/0.92/0.93	0.81/0.86/0.84	0.98/0.97/0.98
LightGBM	0.94/0.92/0.93	0.80/0.88/0.84	0.99/0.97/0.98
ANN	**0.98/0.98/0.98**	**0.92/0.95/0.94**	**0.99/0.99/0.99**

As shown in Table 5, the Moderate risk class consistently produced the lowest F1-scores across all models, ranging from 0.59 (Decision Tree) to 0.94 (ANN). Two factors contribute to this: Moderate is the smallest class (16.39%), and its intermediate position between Low and High creates overlapping feature boundaries with both adjacent classes. This pattern is expected in ordinal psyc

hometric classification, because Moderate risk lies between two adjacent severity regions and may share response profiles with both Low and High classes. Therefore, Moderate-class performance is a more informative stress test of model discrimination than Low–High separation alone. Madububambachu et al. [41] reported the same pattern, finding that intermediate severity categories yielded lower performance across multiple classifiers. The ANN reached a Moderate F1 of 0.94, suggesting that its deeper architecture better captures the non-linear item interactions associated with this risk level [16], [32].

Five of six models produced zero Low-to-High misclassifications, meaning no respondent with low mental health risk was assigned to the high-risk category. Decision Tree was the sole exception, with 5 such misclassifications out of 855 Low-risk test samples (0.6%). However, the more clinically consequential error in this study is High-to-Low underclassification. Therefore, the interpretation of confusion matrices should prioritize whether cost-sensitive learning reduces High-

risk cases incorrectly assigned to Low risk. The rarity of direct Low-to-High errors across five of six models points to strong separation at the extreme ends of the risk spectrum, reducing the likelihood of false alarms that could lead to unnecessary clinical referrals [18].

3.4. Confusion Matrix Analysis

Figure 2 shows the confusion matrices for all six models evaluated on the test set (n = 4,608). The distribution of actual versus predicted risk classes is presented in each matrix.

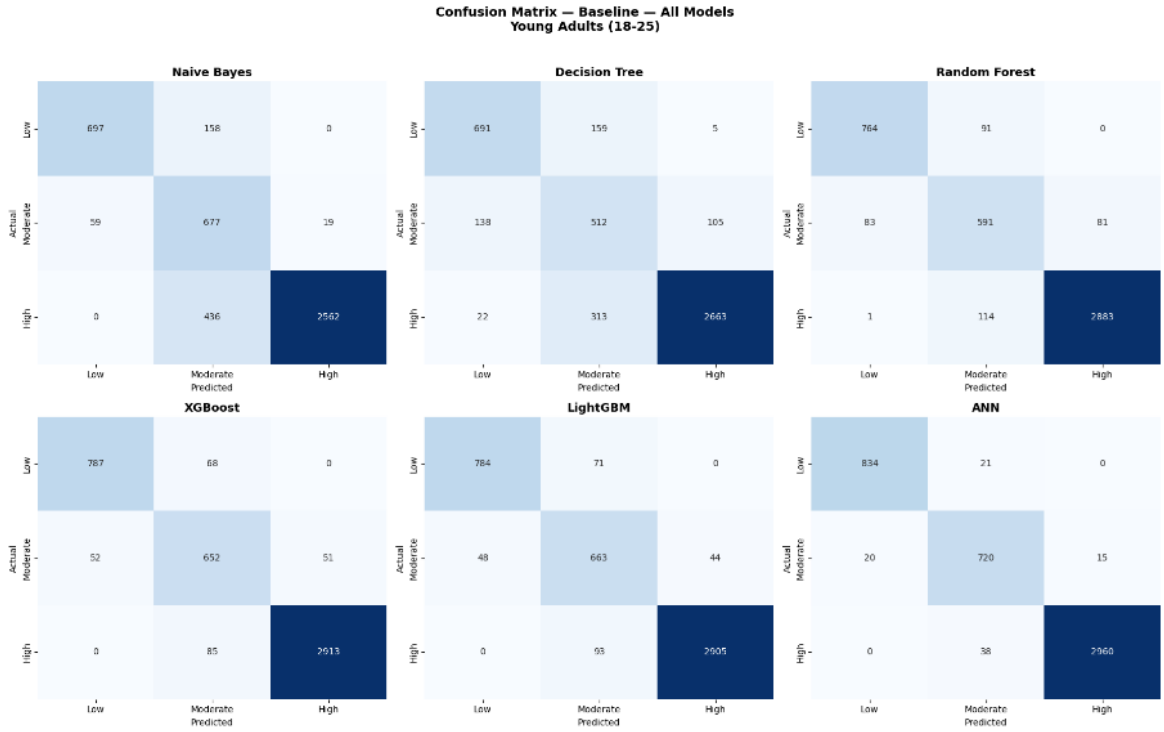


Figure 2. Confusion Matrices Baseline for All Six Models

As shown in Figure 2, the ANN produced the most concentrated diagonal: 21 Low samples were misclassified as Moderate, 20 Moderate as Low, and 15 Moderate as High. Random Forest showed the highest Moderate misclassification counts, with 83 predicted as Low and 81 as High, consistent with its lower Macro F1. LightGBM and XGBoost both produced zero High-to-Low and zero Low-to-High misclassifications. Naive Bayes classified 436 High-risk individuals as Moderate, likely a consequence of its conditional independence assumption, which limits its ability to capture the joint feature interactions in DASS-42 response patterns [27].

3.5. Receiver Operating Characteristic (ROC) Curve Analysis

For every class, the One-vs-Rest (OvR) approach was used to calculate the Receiver Operating Characteristic (ROC) curve with Area Under the Curve (AUC) [38]. The ROC curves for each of the six models are shown in Figure 3, and the per-class AUC values are compiled in Table 6.

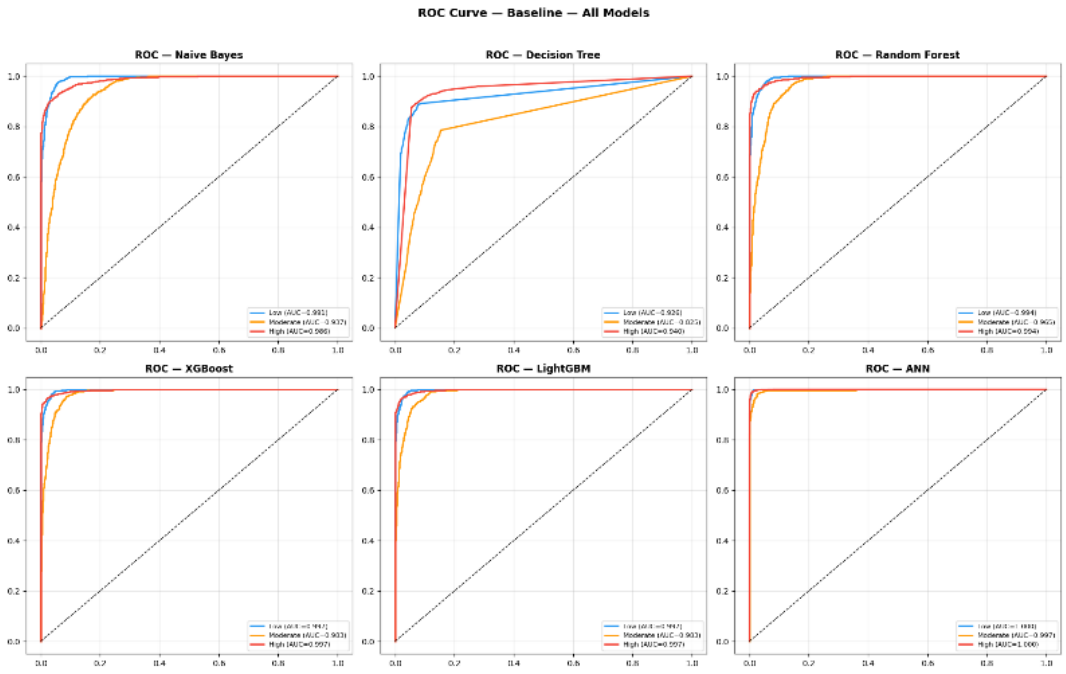


Figure 3. ROC Curves for All Six Models

Table 6. Per-Class and Macro AUC Values

Model	AUC-Low	AUC-Moderate	AUC-High	Macro AUC (avg)
Naive Bayes	0.991	0.937	0.986	0.972
Decision Tree	0.926	0.825	0.940	0.897
Random Forest	0.994	0.965	0.994	0.984
XGBoost	0.997	0.983	0.997	0.992
LightGBM	0.997	0.983	0.997	0.992
ANN	**1.000**	**0.997**	**1.000**	**0.999**

As presented in Table 6, the ANN achieved near-perfect AUC scores across all three classes. XGBoost and LightGBM both reached a macro AUC of 0.992, with particularly strong performance on the High and Low classes. The near-perfect AUC values should be interpreted in the context of DASS-42-derived labels: the model classifies risk categories constructed from the same psychometric item responses, rather than predicting externally confirmed clinical diagnoses. The rule-based deterministic classifier described in Section 2.6 achieves 100% accuracy on this task by construction. Given that baseline, near-perfect AUC and F1 values for the learned models are expected and should not be read as evidence of independent clinical discriminative ability. The Moderate class consistently produced the lowest AUC values across models, with Decision Tree recording the weakest score at 0.825, in line with the per-class F1 results in Table 6. This pattern aligns with findings from Yang et al. [12] and Villanueva-Benito et al. [2] for multi-class mental health risk classifiers on survey-based instruments.

3.6. ANN Training Dynamics

Figure 4 shows the ANN's training and validation accuracy and loss curves under both baseline and CPW configurations.

As shown in Figure 4, the model converged rapidly, reaching above 95% validation accuracy by Epoch 9 and stabilizing near 97 to 99% after Epoch 25. In the early epochs, validation accuracy was higher than training accuracy, which is consistent with dropout being active during training but inhibited during inference [32]. Both loss curves decreased consistently without divergence, suggesting no clear evidence of severe overfitting under the validation split used in this experiment. The best validation accuracy reached 0.9835 for both the baseline and CPW configurations, recorded during EarlyStopping-monitored training. In both cases, EarlyStopping restored the best weights as the final configuration.

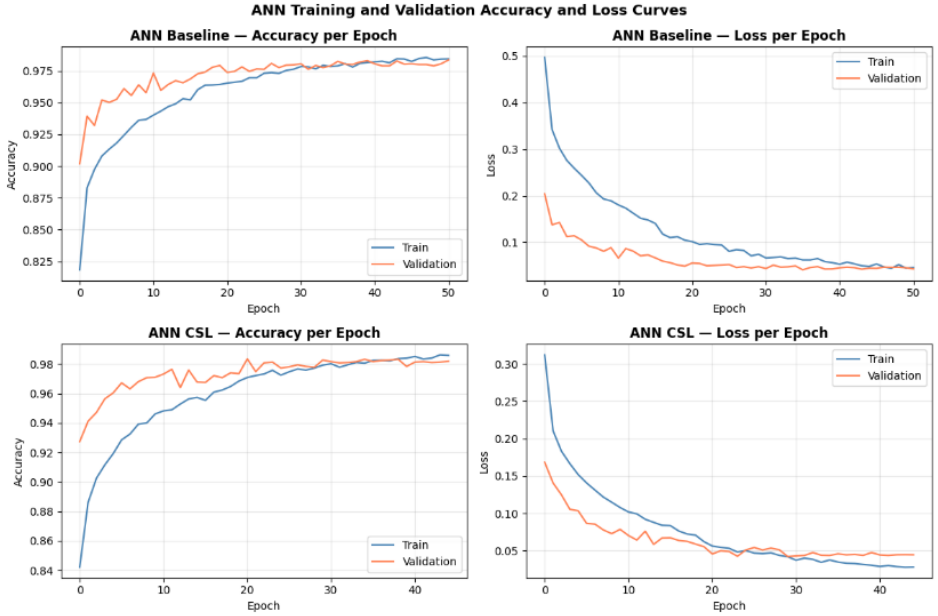


Figure 4. ANN Training and Validation Accuracy and Loss Curves

3.7. Cost-Sensitive Learning Analysis

Macro F1-score alone does not capture the full cost of classification errors in mental health screening. Missing a High-risk individual carries greater clinical consequences than a false positive [17], [18]. To account for this, Cost-Proportionate Weighting (CPW) was applied as described in Section 2, converting the misclassification cost matrix (Table 2) into per-instance training weights [23]. All six models were retrained with CPW-derived sample weights using the same hyperparameter configurations from Section 2. This subsection addresses the second research question by examining whether CPW reduces total misclassification cost relative to the class-balanced baseline. Table 7 compares Macro F1-score and total misclassification cost between the baseline and CPW configurations.

Table 7. Model Performance Comparison: Baseline vs Cost-Sensitive Learning

Model	F1 Baseline	F1 CSL	ΔF1	CV CSL (10-fold)	Total Cost Baseline	Total Cost CSL	Avg Cost Baseline	Avg Cost CSL	ΔCost	Percent Change (%)
Naive Bayes	0.8174	0.8215	+0.0042	0.8345 ± 0.0079	1,603	1,517	0.3479	0.3293	-86	-5.36
Decision Tree	0.7739	0.7745	+0.0006	0.7690 ± 0.0137	1,599	1,381	0.3470	0.2997	−218	-13.63
Random Forest	0.8755	0.8824	+0.0069	0.8769 ± 0.0069	685	549	0.1487	0.1191	-136	**-19.85**
XGBoost	0.9141	0.9100	-0.0041	0.9151± 0.0066	478	402	0.1037	0.0872	-76	-15.90
LightGBM	0.9149	0.9117	-0.0032	0.9206 ± 0.0093	490	454	0.1063	0.0985	-36	-7.35
ANN	0.9686	0.9716	+0.0030	0.9669 ± 0.0040	190	131	0.0412	0.0284	-59	-31.05

As Table 7 shows, CPW reduced total misclassification cost across all six models, though the magnitude varied substantially by algorithm. The ANN showed the largest relative reduction, -31.05% from 190 to 131, followed by Random Forest at -19.85% from 685 to 549, XGBoost at -15.90% from 478 to 402, and Decision Tree at -13.63% from 1,599 to 1,381. Naive Bayes and LightGBM showed smaller but consistent reductions, at -5.36% and -7.35% respectively. The ANN also achieved the highest Macro F1-score under CPW, at 0.9716, and the lowest total misclassification cost among all six models, at 131. Average cost per respondent, computed as total

cost divided by the 4,608 test samples, follows the same pattern as total cost, confirming that the relative benefit of CPW is not an artifact of sample size.

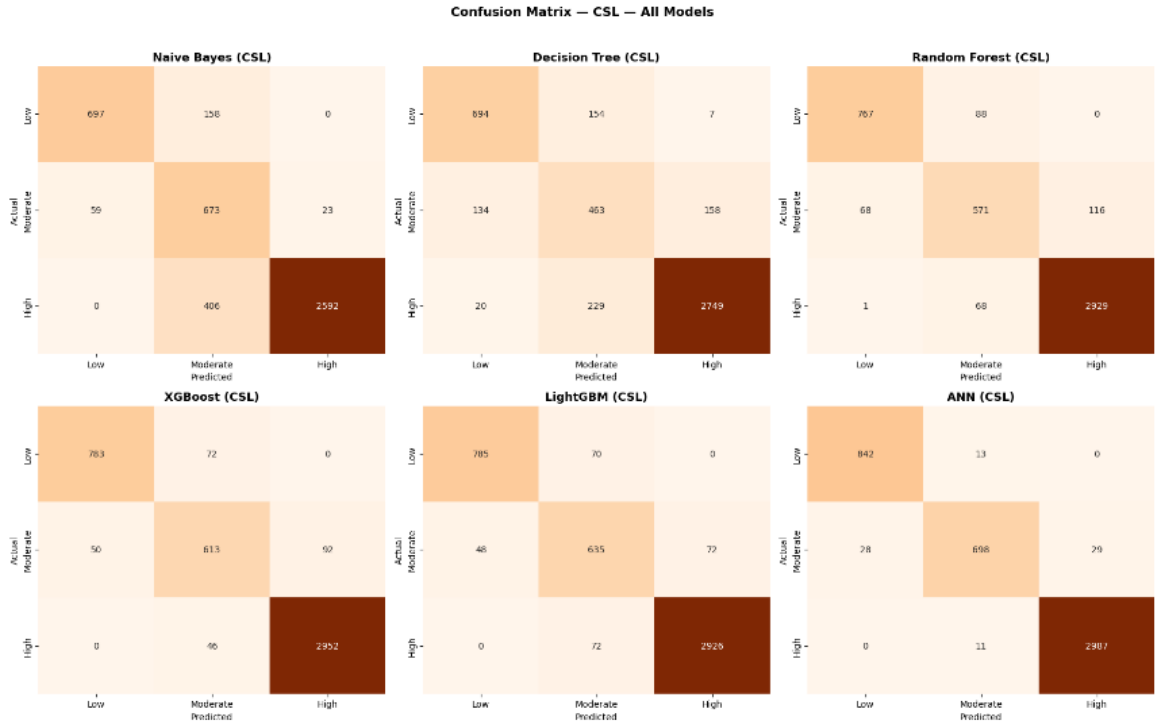


Figure 5. Confusion Matrices CSL for All Six Models

Figure 5 shows how CPW reshapes the error distribution for each model relative to Figure 2. For Random Forest, CPW shifted the balance of Moderate-class errors: Moderate-to-Low misclassifications dropped from 83 to 68, while Moderate-to-High misclassifications rose from 81 to 116. This reflects the higher training weight CPW assigns to the High class, which discourages the model from underclassifying instances as Low at the cost of more conservative High predictions. A similar pattern appears for Naive Bayes, where High-to-Moderate errors fell from 436 to 406, and for Decision Tree, where Low-to-High errors rose slightly from 5 to 7 alongside a drop in High-to-Low errors from 22 to 20. Both patterns match CPW's intended effect: trading a small increase in less costly error types for a reduction in the most clinically consequential High-to-Low errors.

Naive Bayes showed the smallest relative change under CPW (-5.36% cost reduction), despite GaussianNB accepting the *sample_weight* parameter during fitting. A direct comparison confirmed its predictions did differ from the baseline model, which means the earlier characterization of GaussianNB as unaffected by per-instance weighting was imprecise. The classifier does incorporate *sample_weight* into its estimation of class-conditional means and variances, but the resulting shift in decision boundaries was modest relative to tree-based and boosting models in this dataset. This is consistent with its conditional independence assumption, which limits the classifier's overall flexibility to reshape decision boundaries [27]. XGBoost and LightGBM showed cost reductions of -15.90% and -7.35% respectively. Both are smaller than Random Forest's -19.85% and substantially smaller than the ANN's -31.05%, though XGBoost's reduction remains comparable in magnitude to Decision Tree's at -13.63%. This suggests that CPW's benefit does not follow a simple classical-versus-ensemble divide, but instead depends on each model's specific baseline error distribution. These patterns align with findings by Araf et al. [19], who identified that CPW yields the greatest benefit for models with higher baseline susceptibility to majority-class bias, and with Newaz [36] and Jiang et al. [35], who both found cost-sensitive weighting most impactful for simpler classifier architectures. Random Forest and Decision Tree's larger relative gains reflect that pattern, though XGBoost does not fit it: its comparably large gain suggests boosting architecture alone does not determine CPW responsiveness. These results suggest that CPW's effectiveness depends on

whether a model can productively use instance-level cost information and on its baseline decision boundary, rather than on model family alone.

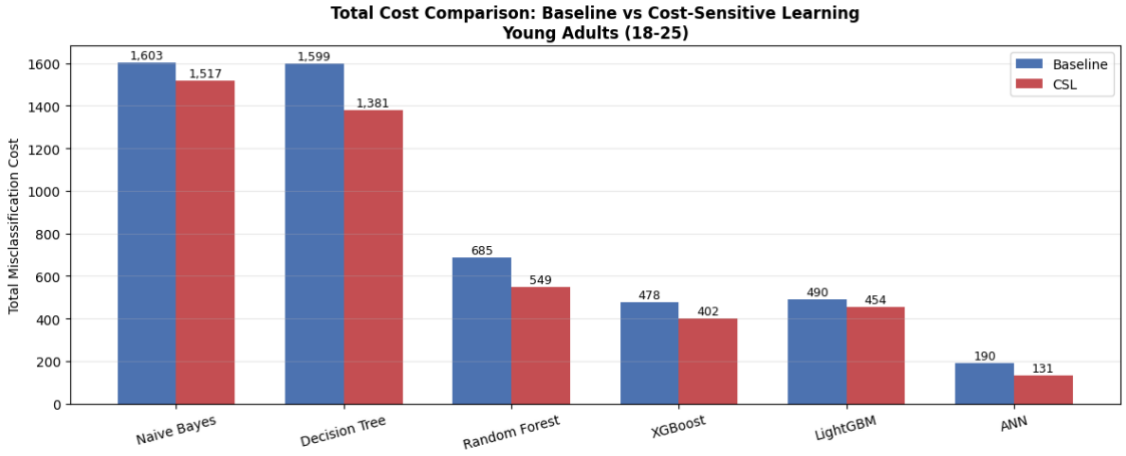


Figure 6. Total Cost Comparison

3.8. Discussion

3.8.1 Research Gap and Positioning Relative to Prior Work

This study was directly motivated by two explicit future research recommendations stated by Biswas et al. [14], whose work represents the most closely related prior study. First, they called for mental health classification on specific at-risk demographic groups rather than the general population. Studies using the same DASS-42 Kaggle dataset, including ShamsEldin et al. [42], did not apply age-based filtering, and Park and Woo [43] similarly noted that young adults require targeted analytical frameworks distinct from general-population models. Second, Biswas et al. recommended more robust statistical evaluation through cross-validation. The present study addresses both recommendations: the dataset was restricted to adults aged 18 to 25, a group with documented elevated rates of psychological distress [3], [26], and 10-fold Stratified K-Fold Cross-Validation was applied across all six models.

A third contribution extends beyond these recommendations: the integration of CPW as a cost-sensitive learning strategy [23], which was absent in Biswas et al. [14]. Together, these three methodological choices distinguish this study from the broader DASS-42 machine learning literature [11], [12], [40]. Each addresses a gap identified in recent systematic reviews. Wajid et al. [7] flagged the absence of age-stratified analyses; Chahar and Mishra [8] noted the prevalence of single-split evaluations without cross-validation. Hida et al. [24] identified cost-sensitive learning as an underused approach for imbalanced health classification datasets, while Razavi et al. [44] similarly identified the lack of standardized data preprocessing protocols as a persistent gap across stress-related ML research, a concern addressed here through the explicit, reproducible preprocessing pipeline described in Section 2.2-2.3.

A fourth contribution addresses a methodological gap not previously raised in the DASS-42 literature: the explicit testing and disclosure of label-feature circularity. Biswas et al. [14], Kumar et al. [11], and other DASS-42 studies construct risk labels directly from the same items used as predictors, so their reported near-perfect performance metrics are subject to the same circularity documented here (Section 2.6, Section 3.8.6). This study is, to our knowledge, the first in the DASS-42 machine learning literature to quantify this circularity through an explicit rule-based baseline, and to reframe its contribution accordingly rather than report predictive accuracy without that caveat.

3.8.2 Comparison with Prior Work

Table 8 presents a systematic comparison of the present study with key prior works in DASS-based and related mental health machine learning classification.

Table 8. Comparison with Prior Studies on ML-Based Mental Health Classification

Study	Dataset/Population	Models	Best Model	Best Metric	Differences
Biswas et al. [14]	DASS-42 Kaggle, general population	KNN, RF, SVM, Ensemble, ANN	ANN	Acc: 99.73% (depression)	No age stratification, no cost sensitive analysis
Kumar et al. [11]	DASS-42, general population	SVM, RF, KNN, DT, NB	RF	F1: 0.87 (multi-class)	No age stratification, no cost sensitive analysis
Yang et al. [12]	Anxiety survey, general population	XGBoost, LightGBM, RF, ANN	LightGBM	AUC: 0.94	Focus only on anxiety, no age stratification, no cost sensitive analysis
Bader et al. [17]	PHQ-9 depression, adults, CPW	RF, WRF, LR, WLR	WRF+CPW	AUC: 0.91 (multi-class)	Focus only on depression
This Study	DASS-42, young adults 18–25, CPW	NB, DT, RF, XGBoost, LightGBM, ANN	ANN+CPW	F1: 0.9716 (CSL), AUC: 0.999 (baseline), TC: 131 (CSL, ↓31.05% vs baseline)	Age specific, asymmetric cost sensitive analysis, explicit acknowledgment of label-feature circularity via rule-based deterministic baseline

As shown in Table 8, to the best of our knowledge, prior DASS-42 studies have not jointly combined demographic specificity (adults aged 18 to 25), multi-class global risk classification, 10-fold cross-validation, and cost-sensitive learning within a single framework. The ANN achieved a Macro F1 of 0.9716 under CPW, comparable to ANN-based benchmarks in the literature. The inclusion of total misclassification cost (TC = 131) as a complementary metric adds a clinically interpretable dimension that prior DASS-42 studies do not report [11], [12], [14].

Direct numerical comparison with Biswas et al. [14] requires care: their study classified each domain independently into five severity levels, used an unfiltered dataset, and applied Random Oversampling rather than class weight balancing. The lower absolute accuracy in the present study reflects more constrained training data, a more homogeneous target population, and a stricter evaluation protocol. Ahammed et al. [3] observed a similar pattern, noting that models trained on age-stratified youth cohorts tend to show greater prediction uncertainty than general-population models.

3.8.3 The Moderate Class Challenge

Across all six models, the Moderate risk class produced the lowest F1-scores (0.59 to 0.94) and the lowest ROC-AUC values (0.825 to 0.997). Two factors explain this. The Moderate class is the smallest in the dataset (n = 3,776; 16.39%), so models have fewer examples from which to learn discriminative patterns. It also occupies an intermediate position where DASS-42 response patterns overlap with both adjacent classes. This overlap is not a statistical artifact; it reflects genuine clinical ambiguity in the intermediate range of psychological distress, documented in the DASS-42 validation literature [20]. For screening practice, this means that future work should not only optimize overall Macro F1-score but also examine calibration and decision thresholds around the Moderate–High boundary, where intervention decisions are most sensitive.

Madububambachu et al. [41] reported the same pattern in their systematic review, finding that intermediate severity categories consistently yield lower classification performance across multiple algorithms. The ANN reached a Moderate F1 of 0.94, compared to 0.67 for Naive Bayes and 0.84 for both XGBoost and LightGBM, suggesting that deeper non-linear architectures are better suited to resolving ambiguity in this class. From a clinical standpoint, accurate identification of Moderate-risk individuals is particularly important for early intervention [16], [18], as this group occupies the threshold between requiring and not yet requiring formal mental health support.

3.8.4 Interpretation of Extreme-Class Misclassification Patterns

Five of six models produced zero Low-to-High misclassifications across the test set of 4,608 samples; Decision Tree was the sole exception (Section 3.3). For the five models without this error type, no DASS-42-defined low-risk individual was assigned to the high-risk category. This finding is practically relevant for screening support because such over-prediction errors, while less dangerous than false negatives, can lead to unnecessary referrals and inefficient allocation of limited mental health resources [19], [37].

This outcome suggests that the 42-item DASS-42 response profile creates sufficient feature separation between Low and High risk for all tested models to learn reliably, including Naive Bayes. The result is consistent with the DASS-42 design principle, which holds that normal and clinical-range psychological distress differ dimensionally rather than categorically, making extreme profiles at either end of the severity spectrum more distinctive than intermediate ones.

Dino et al. [18] noted in their systematic review that false negatives on high-severity classes are the most clinically consequential error type in mental health classification. Five of six models produced neither extreme error type (Low to High nor High to Low) under either the baseline or CPW configuration, indicating promising extreme-class separation within this DASS-42-derived screening task, consistent with the growing emphasis on early-stage risk assessment frameworks for student populations [9]. Decision Tree was the sole exception, producing both error types under both configurations: Low-to-High errors rose from 5 under baseline to 7 under CSL, while High-to-Low errors fell from 22 to 20. This is consistent with its comparatively lower overall performance (Table 4). This result aligns with the strong overall performance reported by Biswas et al. [14] on the same dataset, where ANN achieved near-perfect accuracy for DASS-42 classification tasks.

3.8.5 Contribution and Limitations of Cost-Proportionate Weighting

As detailed in Section 3.7, CPW's benefit varied substantially by model architecture, with the ANN and Random Forest showing the largest relative reductions and Naive Bayes the smallest. This pattern is broadly consistent with the cost-sensitive learning literature. Araf et al. [19] found that CSL improvements tend to be larger for models more susceptible to majority-class bias, and Newaz [36] and Jiang et al. [35] similarly reported the largest cost-sensitive gains for simpler classifier architectures, though the ANN's disproportionately large reduction here suggests factors beyond model simplicity also drive CPW responsiveness (Section 3.8.6). The most directly analogous prior application of CPW to a mental health problem is Bader et al. [17], who applied it to Weighted Random Forest for depression severity classification and reported an AUC improvement from 0.88 to 0.91. The TC metric introduced here offers a practical, clinically motivated basis for model selection beyond F1-score alone [19], [33].

The ANN's disproportionately large relative cost reduction, despite being the most complex model evaluated, may reflect its comparatively higher baseline cost per correct-classification rate (Table 7). Because the ANN's baseline error count was small in absolute terms, a modest absolute change in misclassification pattern under CPW produces a large percentage change. This underscores that percentage-based cost reduction should be interpreted alongside absolute cost values (Table 7) rather than in isolation.

3.8.6 Label-Feature Circularity and Reframed Contribution

We emphasize that the deterministic scoring rule, being error-free on this task by construction, also incurs zero misclassification cost. No learned model can improve upon it in either accuracy or cost for the exact task of reconstructing DASS-42-derived labels from DASS-42 items. We therefore withdraw any claim that the learned models add practical value over the deterministic rule for this specific reconstruction task. Instead, this study should be understood as a controlled methodological case study. Because the ground-truth decision boundary is known exactly, this setting allows direct observation of how cost-sensitive training (CPW) reshapes model error distributions relative to a known-optimal reference point, an evaluation not possible in real-world screening settings where ground truth is uncertain.

Beyond this diagnostic value, we identify one concrete capability the deterministic rule structurally cannot provide: continuous, cross-domain-informed confidence estimates. The rule assigns each respondent a discrete label with no indication of proximity to a decision boundary. A respondent scoring one point above a threshold is treated identically to one scoring twenty points

above it, and the rule considers only the single highest-severity domain, ignoring information from the other two. The learned models, in contrast, produce class probabilities that vary continuously and may integrate information across all three domains. To test whether this confidence reflects more than a restatement of the rule's own single-domain margin, we computed, for each test respondent, a rule-based margin, defined as the distance between the domain score determining the respondent's severity classification and its nearest applicable threshold, and correlated it with the predicted-class probability of all six models using Spearman rank correlation.

The correlation was substantial but clearly below the deterministic bound of 1.0 for every model (Naive Bayes $\rho = 0.782$; Decision Tree $\rho = 0.464$; Random Forest $\rho = 0.877$; XGBoost $\rho = 0.888$; LightGBM $\rho = 0.916$; ANN $\rho = 0.803$; all $p < 0.001$), indicating this pattern is not specific to any single architecture. Notably, Decision Tree showed the weakest correlation with the rule margin, consistent with its discrete, threshold-based decision structure, which diverges most from a continuous margin. LightGBM and XGBoost showed the strongest correlations among the six models, suggesting gradient-boosted trees track the single-domain margin more closely than the ANN, even though the ANN's confidence estimates diverge from that margin more than these two. This indicates that model confidence is related to, but not fully determined by, the univariate margin the rule relies on, suggesting the learned models partly integrate cross-domain information. A respondent borderline on the Stress domain but clearly elevated on Depression, for instance, may receive a different confidence estimate than the single-domain margin alone would predict, a nuance the maximum-severity rule cannot represent by design. We note that this margin operationalization is domain-specific and not perfectly comparable in scale across depression, anxiety, and stress. We therefore treat this analysis as an exploratory illustration of a potential confidence-based triage use case rather than a definitive validated application, and recommend that future work formalize and validate this triage capability on prospectively collected data.

4. CONCLUSION

This study addressed three methodological gaps in DASS-42-based mental health risk classification: limited age-specific analysis, insufficient validation robustness, and the lack of cost-sensitive evaluation for clinically unequal screening errors. Using 23,039 respondents aged 18–25, this study evaluated six machine learning models for three-class DASS-42 risk classification under class-balanced and cost-sensitive configurations. The proposed framework should be interpreted as automated DASS-42-based screening support rather than prediction of externally confirmed psychiatric diagnoses, because the target labels were derived from DASS-42 item responses.

The findings answer the two research questions. First, the Artificial Neural Network achieved the strongest held-out performance, with the highest Macro F1-score and macro AUC, whereas XGBoost showed the most stable cross-validation performance among the classical models. Second, Cost-Proportionate Weighting reduced total misclassification cost across all six models, with reductions ranging from 5.4% to 31.05%. The ANN achieved both the largest relative cost reduction and the lowest final total misclassification cost among all models. These results show that cost-sensitive weighting is useful but not universally beneficial across all classifiers; its effectiveness depends on model architecture and baseline error patterns.

Overall, this study makes two main contributions. Methodologically, it offers a controlled diagnostic case study of how cost-sensitive weighting reshapes model error distributions relative to a known-optimal, error-free reference point. Practically, it demonstrates that learned models can produce continuous, cross-domain-informed confidence estimates for borderline-case triage, a capability the deterministic DASS-42 scoring rule cannot offer by design, even though no learned model can outperform it in accuracy or cost on the label-reconstruction task itself. Nevertheless, the findings remain limited by three factors: reliance on a secondary self-report dataset, a structural circularity between the DASS-derived target labels and the item-level features used to predict them (Section 3.8.6), and a cost matrix that is clinically motivated rather than empirically calibrated. Future work should validate the framework on prospectively collected and clinically verified data, calibrate the cost matrix through consultation with a clinical expert panel or empirical health economic evidence rather than the authors' reasoning alone, and apply explainability methods such as SHAP to identify which DASS-42 items contribute most to each risk classification. Incorporating additional

metadata and response validity fields available in the original survey, such as response duration and attention check indicators, may further improve data quality screening prior to model training. These steps are necessary before the framework can be considered for implementation as an internet delivered or mobile screening support tool in university mental health programs.

ACKNOWLEDGEMENTS
This research received no external funding and was conducted independently by the authors.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

DATA AVAILABILITY
The data associated with this study are publicly available online in the replication package. https://bit.ly/4wjNDlh

AUTHOR CONTRIBUTIONS
Iklima Fatma Aliena: Conceptualization; Methodology; Software; Formal analysis; Investigation; Writing-Original Draft; Visualization. **Riska Dhenabayu:** Conceptualization; Methodology; Supervision; Writing-Review & Editing; Resources. **Rully Mujiastuti:** Conceptualization; Methodology; Writing - Original Draft; Funding Acquisition. **Tatiana Starostina:** Validation; Project administration. **Ratih Amelia:** Validation; Project administration.

REFERENCES
[1] World Health Organization, "Mental disorders." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] G. V. Benito, "Machine learning for anxiety and depression profiling and risk assessment in the aftermath of an emergency," *Artif Intell Med*, vol. 157, p. 102991, 2024, doi: 10.1016/j.artmed.2024.102991.

[3] B. Ahammed, K. Alam, Z. Hoque, and S. A. Keramat, "Mental health at risk: Predicting psychological distress in Australian youth through machine learning models," *J. Affect. Disord.*, vol. 395, p. 120679, Feb. 2026, doi: 10.1016/j.jad.2025.120679.

[4] W. Li, Z. Zhao, D. Chen, Y. Peng, and Z. Lu, "Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis," *J. Child Psychol. Psychiatry*, vol. 63, no. 11, pp. 1222–1230, Nov. 2022, doi: 10.1111/jcpp.13606.

[5] A. Daza, N. Saboya, J. I. Necochea-Chamorro, K. Zavaleta Ramos, and Y. D. R. Vásquez Valencia, "Systematic review of machine learning techniques to predict anxiety and stress in college students," *Inform. Med. Unlocked*, vol. 43, p. 101391, 2023, doi: 10.1016/j.imu.2023.101391.

[6] R. Zhao *et al.*, "Research Review: Help-seeking intentions, behaviors, and barriers in college students – a systematic review and meta-analysis," *J. Child Psychol. Psychiatry*, vol. 66, no. 10, pp. 1593–1605, Oct. 2025, doi: 10.1111/jcpp.14145.

[7] A. Wajid, F. Azam, and M. W. Anwar, "Applications of artificial intelligence in mental health: a systematic literature review," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 332, Nov. 2025, doi: 10.1007/s44163-025-00569-2.

[8] R. Chahar and A. K. Dubey, "Machine Intelligence for Mental Health Diagnosis: A Systematic Review of Methods, Algorithms, and Key Challenges," *Comput. Mater. Contin.*, vol. 86, no. 1, pp. 1–65, 2026, doi: 10.32604/cmc.2025.066990.

[9] X. Shi, J. Pan, D. Yuan, M. Li, and Y. Pan, "Advanced data analysis and prediction model for student mental health risk assessment," *Front. Psychol.*, vol. 16, p. 1682083, Nov. 2025, doi: 10.3389/fpsyg.2025.1682083.

[10] M. Makara-Studzińska, E. Tyburski, M. Załuski, K. Adamczyk, J. Mesterhazy, and A. Mesterhazy, "Confirmatory Factor Analysis of Three Versions of the Depression Anxiety Stress Scale (DASS-42, DASS-21, and DASS-12) in Polish Adults," *Front. Psychiatry*, vol. 12, p. 770532, Jan. 2022, doi: 10.3389/fpsyt.2021.770532.

[11] P. Kumar, S. Garg, and A. Garg, "Assessment of Anxiety, Depression and Stress using Machine Learning Models," *Procedia Comput. Sci.*, vol. 171, pp. 1989–1998, 2020, doi: 10.1016/j.procs.2020.04.213.

[12] H. Yang, Y. H. Sun, and K. Lee, "Concise multi-class anxiety disorder risk assessment: A novel advanced machine learning approach," *J. Anxiety Disord.*, vol. 112, p. 103018, Jun. 2025, doi: 10.1016/j.janxdis.2025.103018.
[13] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms," *Healthc. Anal.*, vol. 3, p. 100185, 2023, doi: https://doi.org/10.1016/j.health.2023.100185.
[14] J. Biswas, Md. N. Hasan, Md. S. R. Gazi, and Md. M. Rahman, "Enhancing mental well-being: An artificial intelligence model for predicting mental disorders," *Array*, vol. 26, p. 100417, Jul. 2025, doi: 10.1016/j.array.2025.100417.
[15] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," *Healthcare*, vol. 11, no. 3, p. 285, Jan. 2023, doi: 10.3390/healthcare11030285.
[16] R. Dehbozorgi *et al.*, "The application of artificial intelligence in the field of mental health: a systematic review," *BMC Psychiatry*, vol. 25, no. 1, p. 132, Feb. 2025, doi: 10.1186/s12888-025-06483-2.
[17] M. Bader, M. Abdelwanis, M. Maalouf, and H. F. Jelinek, "Detecting depression severity using weighted random forest and oxidative stress biomarkers," *Sci. Rep.*, vol. 14, no. 1, p. 16328, Jul. 2024, doi: 10.1038/s41598-024-67251-y.
[18] H. I. Dino and M. M. Hassan, "Classification of neurological and mental health disorders based on multimodal approaches: A comprehensive review," *Neurosci. Biobehav. Rev.*, vol. 179, p. 106399, Dec. 2025, doi: 10.1016/j.neubiorev.2025.106399.
[19] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev.*, vol. 57, no. 4, p. 80, Mar. 2024, doi: 10.1007/s10462-023-10652-8.
[20] S. H. Lovibond and P. F. Lovibond, *Manual for the depression anxiety stress scales*, 2. ed. in Psychology Foundation monograph. Sydney: Psychology Foundation, 1995.
[21] L. Sasse *et al.*, "Overview of leakage scenarios in supervised machine learning," *J. Big Data*, vol. 12, no. 1, p. 135, May 2025, doi: 10.1186/s40537-025-01193-8.
[22] M. Hagmann and S. Riezler, "False perfection in machine prediction: Detecting and assessing circularity problems in machine learning," Dec. 13, 2021, *arXiv*: arXiv:2106.12417. doi: 10.48550/arXiv.2106.12417.
[23] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Third IEEE International Conference on Data Mining*, 2003, pp. 435–442. doi: 10.1109/ICDM.2003.1250950.
[24] G. S. Hida and A. C. Alves Do Nascimento, "Overview of machine learning in class imbalance scenarios: Trends, challenges, and approaches," *Expert Syst. Appl.*, vol. 298, p. 129592, Mar. 2026, doi: 10.1016/j.eswa.2025.129592.
[25] L. Greenwell, "Depression Anxiety Stress Scales Responses." 2019. [Online]. Available: https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses
[26] Z. Wang *et al.*, "Global, regional, and national burden of mental disorders among adolescents and young adults, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Transl. Psychiatry*, vol. 15, no. 1, p. 397, Oct. 2025, doi: 10.1038/s41398-025-03623-w.
[27] I. Rish, "An empirical study of the naive Bayes classifier," 2001.
[28] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Boca Raton London New York Washington, D.C: Chapman & Hall/CRC, 1984. doi: 10.1201/9781315139470.
[29] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[30] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[31] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," 2017.
[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[33] B. Davazdahemami, P. Peng, and D. Delen, "A deep learning approach for predicting early bounce-backs to the emergency departments," *Healthc. Anal.*, vol. 2, p. 100018, Nov. 2022, doi: 10.1016/j.health.2022.100018.
[34] B. Davazdahemami, H. M. Zolbanin, and D. Delen, "An explanatory machine learning framework for studying pandemics: The case of COVID-19 emergency department readmissions," *Decis. Support Syst.*, vol. 161, p. 113730, Oct. 2022, doi: 10.1016/j.dss.2022.113730.

[35] Z. Jiang and W. Zhao, "Fusion Algorithm for Imbalanced EEG Data Processing in Seizure Detection," *Seizure*, vol. 91, pp. 207–211, Oct. 2021, doi: 10.1016/j.seizure.2021.06.023.
[36] A. Newaz, M. S. Mohosheu, and Md. A. Al Noman, "Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques," *Inform. Med. Unlocked*, vol. 42, p. 101361, 2023, doi: 10.1016/j.imu.2023.101361.
[37] L. Rabkin, I. Cohen, and G. Singer, "Resource allocation in ordinal classification problems: A prescriptive framework utilizing machine learning and mathematical programming," *Eng. Appl. Artif. Intell.*, vol. 132, p. 107914, Jun. 2024, doi: 10.1016/j.engappai.2024.107914.
[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[39] W. L. Ku and H. Min, "Evaluating Machine Learning Stability in Predicting Depression and Anxiety Amidst Subjective Response Errors," *Healthcare*, vol. 12, no. 6, p. 625, Mar. 2024, doi: 10.3390/healthcare12060625.
[40] Md. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, "An in-depth analysis of machine learning approaches to predict depression," *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, Nov. 2021, doi: 10.1016/j.crbeha.2021.100044.
[41] U. Madububambachu, A. Ukpebor, and U. Ihezue, "Machine Learning Techniques to Predict Mental Health Diagnoses: A Systematic Literature Review," *Clin. Pract. Epidemiol. Ment. Health*, vol. 20, no. 1, p. e17450179315688, Jul. 2024, doi: 10.2174/0117450179315688240607052117.
[42] T. ShamsEldin *et al.*, "Artificial intelligence for predicting depression anxiety and stress using psychometric data," *Sci. Rep.*, vol. 15, no. 1, p. 37282, Oct. 2025, doi: 10.1038/s41598-025-21301-1.
[43] N. Park and H. Woo, "Change in lifestyle and mental health in young adults: an exploratory study with hybrid machine learning," *Front. Public Health*, vol. 13, p. 1562280, Jun. 2025, doi: 10.3389/fpubh.2025.1562280.
[44] M. Razavi *et al.*, "Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review," *JMIR Ment. Health*, vol. 11, p. e53714, Aug. 2024, doi: 10.2196/53714.