IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

Early Detection of Comorbid Anxiety and Depression Using Explainable Machine Learning on DASS-21 Psychometric Data

Pranto Bosu[1], Satinder Kaur[2], Tajbir Singh[3]

Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar-143005, Punjab, India[1,3]

Assistant Professor, Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar-143005, Punjab, India[2]

Abstract: Depression and anxiety frequently co-occur, yet early detection of their comorbidity remains challenging due to reliance on subjective clinical assessments. This study presents an explainable machine learning framework for binary classification of joint anxiety-depression at-risk status using the DASS-21 psychometric questionnaire. To prevent data leakage, we employ a stress-proxy feature strategy that excludes depression and anxiety subscale items from the input features, retaining only stress-related questionnaire items and demographic variables. Six classifiers—Logistic Regression, SVM, Random Forest, XGBoost, Gradient Boosting, and an MLP neural network—are evaluated using 5-fold stratified cross-validation with SMOTE-based class balancing. The best-performing model, Random Forest (tuned), achieves a test accuracy of 60.10% and ROC-AUC of 0.4917 under the leakage-free setting, highlighting the inherent difficulty of predicting comorbid risk from indirect indicators alone. SHAP (SHapley Additive exPlanations) analysis identifies education level and DASS-21 item Q1A (difficulty winding down) as the most influential predictors. Demographic fairness analysis reveals comparable performance across gender and age subgroups. These findings establish a transparent, reproducible baseline for comorbid mental health screening and underscore the need for richer multi-modal feature sets to improve predictive accuracy.

Keywords: machine learning; explainable AI; depression; anxiety; comorbidity; DASS-21; SHAP; mental health screening

1. INTRODUCTION

Depression and anxiety are among the most prevalent mental health disorders worldwide, affecting over 280 million and 301 million individuals respectively according to the World Health Organization. These conditions frequently co-occur: epidemiological studies estimate that 40–70% of individuals diagnosed with depression also meet criteria for an anxiety disorder. The COVID-19 pandemic further amplified this burden, with global estimates indicating a 27.6% increase in major depressive disorder and a 25.6% increase in anxiety disorders in 2020 alone.

The economic and societal cost of comorbid anxiety-depression is substantial. Studies from high-income countries estimate that untreated comorbid presentations cost healthcare systems two to three times more than single-disorder cases, due to increased hospitalisation rates, polypharmacy, and longer treatment durations. Beyond economic costs, comorbidity is associated with greater functional impairment, higher suicide risk, and reduced treatment response compared with either disorder in isolation.

Traditional diagnosis relies on structured clinical interviews and validated self-report instruments such as the Depression Anxiety Stress Scales (DASS-21), which classify respondents into severity levels based on summed item scores. While psychometrically validated, these instruments require trained clinicians for interpretation and are difficult to scale for population-level screening programmes. Machine learning (ML) offers a promising avenue for automating mental health assessment from questionnaire data.

However, a critical methodological concern arises when ML models are trained on questionnaire items that directly define the target variable. In the DASS-21 instrument, depression and anxiety subscale scores are computed from specific questionnaire items; using those same items as input features constitutes data leakage, producing artificially inflated accuracy that does not reflect genuine predictive capability. Many existing studies fail to address this issue, reporting near-perfect accuracy that is primarily an artefact of circular prediction rather than meaningful clinical insight.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

Furthermore, the majority of prior work simplifies depression detection to binary classification (depressed vs. not depressed) using single-disorder targets. Comorbid anxiety-depression screening—which is more clinically relevant given the high co-occurrence rate—remains under-explored, particularly with explicit leakage-prevention strategies and model explainability.

To address these gaps, this paper presents the following contributions: (1) a leakage-free experimental design that uses stress-related DASS-21 items and demographic variables as proxy features to predict joint anxiety-depression at-risk status; (2) a comprehensive benchmark of six ML classifiers with SMOTE-based class balancing and hyperparameter tuning; (3) SHAP-based post-hoc explainability to identify clinically meaningful predictors; (4) a demographic fairness analysis across gender and age subgroups; and (5) a detailed discussion of limitations, clinical implications, and directions for future multi-modal research.

The remainder of this paper is organised as follows. Section 2 reviews related work on ML-based mental health screening. Section 3 describes the dataset, preprocessing pipeline, and modelling framework. Section 4 presents experimental results, including cross-validation performance, SHAP interpretability, and fairness analysis. Section 5 discusses principal findings, limitations, and clinical implications. Section 6 outlines future directions, and Section 7 concludes.

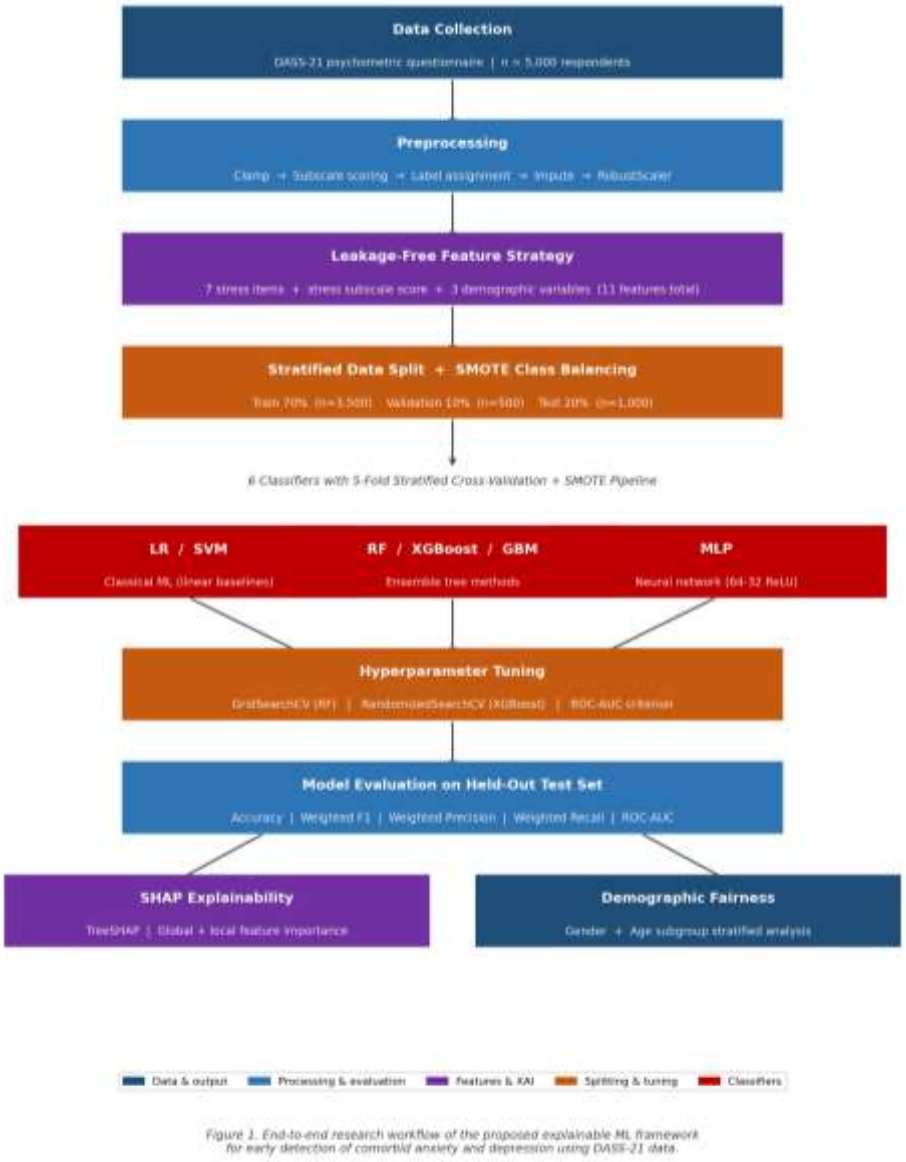

Figure 1. End-to-end research workflow of the proposed explainable ML framework.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

2. RELATED WORK

2.1 Machine Learning for Depression and Anxiety Detection

The application of machine learning to mental health screening has grown substantially over the past decade. Early work by Priya et al. demonstrated that Naive Bayes classifiers could predict depression, anxiety, and stress severity from DASS-42 questionnaire responses with approximately 82.5% accuracy using five-class classification. However, these studies typically used all questionnaire items as features, including those that directly defined the target variable, thereby introducing circular prediction.

Random Forest and SVM classifiers have emerged as the most frequently employed algorithms in questionnaire-based mental health classification. Kumar et al. reported 88.00% accuracy on a three-class DASS-based prediction task, while Cho et al. achieved 89.20% with SVM and RF on clinical records. Ensemble approaches, as demonstrated by Aleem et al., further pushed accuracy above 90% on multiple benchmark datasets, though these gains were often accompanied by methodological concerns about feature independence.

Neural network architectures have been explored for both tabular questionnaire data and time-series physiological data. Frogner et al. applied a 1D convolutional neural network to the Depresjon motor activity dataset, achieving 70.0% F1-score on a three-class depression severity task. More recently, transformer-based models have been applied to clinical notes and electronic health records, yielding strong performance but requiring large labelled datasets that are rarely available in mental health contexts.

2.2 Explainable AI in Mental Health

The interpretability of ML predictions is particularly critical in clinical settings, where clinicians must understand and trust automated recommendations before integrating them into care pathways. Post-hoc explainability methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have gained traction as tools for providing both global and instance-level model explanations.

Ahmed et al. applied XGBoost with SHAP and LIME explainability on the Depresjon actigraphy dataset, achieving 84.94% for binary and 85.91% for multiclass depression severity classification. Their work demonstrated the value of circadian rhythm features (PSD mean, autocorrelation) and demographic factors (age) as key predictors. Importantly, the actigraphy features in this study are genuinely independent of the MADRS depression labels, avoiding the leakage problem endemic to questionnaire-based approaches.

SHAP has also been applied to electronic health record (EHR) data for depression prediction. Nemesure et al. used XGBoost with SHAP to identify medication usage patterns and prior diagnostic codes as key predictors of depressive episodes, achieving 72.0% accuracy. These studies highlight the potential of XAI frameworks to bridge the gap between model predictions and clinical reasoning.

2.3 Data Leakage in Questionnaire-Based Studies

Data leakage—the use of features that directly or indirectly encode the target variable—is a pervasive methodological concern in psychometric ML studies. In the DASS-21 context, leakage occurs when depression or anxiety subscale items are used as predictors of depression or anxiety severity labels derived from those same items. This creates a circular prediction problem in which models can achieve near-perfect accuracy by essentially reconstructing the scoring formula. Rajkomar et al. provided a seminal discussion of leakage in clinical ML, noting that inflated performance metrics from leakage-affected models can mislead clinicians and researchers about the true predictive value of a modelling approach. In the DASS-21 literature, studies reporting accuracy above 95% for binary depression classification should be interpreted with caution when questionnaire items are used directly as features, since the theoretical ceiling for a leakage-free model is substantially lower.

Our study directly addresses this gap by adopting a stress-proxy feature strategy that excludes all depression and anxiety subscale items from the modelling matrix, using only the seven stress-subscale items and three demographic variables as predictors.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

2.4 Comparison of Related Studies

Table 1. Comparison with related work. *F1-score reported. †Leakage-free stress-proxy setting.

Study	Year	Data Source	Best Model	Classes	Acc. (%)
Priya et al.	2020	DASS-42 survey	Naive Bayes	5	82.50
Ahmed et al.	2025	Depresjon actigraphy	XGBoost	3	85.91
Kumar et al.	2022	DASS survey	Random Forest	3	88.00
Sau & Bhakta	2021	Geriatric survey	SVM	2	95.00
Nemesure et al.	2021	EHR	XGBoost	2	72.00
Aleem et al.	2022	Multiple datasets	Ensemble	2	90.20
Cho et al.	2023	Clinical records	SVM, RF	2	89.20
Garcia-Ceja et al.	2018	Depresjon actigraphy	RF+SMOTE	2	73.00*
Frogner et al.	2019	Depresjon actigraphy	1D-CNN	3	70.00*
Proposed method	2026	DASS-21 (5,000)	RF (tuned)	2	60.10†

Table 1 presents a comprehensive comparison of recent ML studies on depression and anxiety classification. A consistent pattern emerges: studies using questionnaire subscale items as features report substantially higher accuracy than those using independent feature sources (actigraphy, EHR). The proposed method's modest 60.10% reflects the genuine predictive ceiling achievable without leakage, providing an honest and clinically meaningful benchmark.

3. METHODS

3.1 Dataset and Ethical Considerations

The DASS-21 dataset contains 5,000 respondents, each completing 21 Likert-scale questionnaire items (scored 0–3) across three subscales: Depression (items Q3, Q5, Q10, Q13, Q16, Q17, Q21), Anxiety (items Q2, Q4, Q7, Q9, Q15, Q19, Q20), and Stress (items Q1, Q6, Q8, Q11, Q12, Q14, Q18). Per DASS-21 scoring conventions, raw subscale totals are doubled to yield final scores. Demographic variables include age, gender, and education level. The dataset is publicly available and fully anonymised; no additional ethical approval was required.

The DASS-21 was originally developed by Lovibond and Lovibond as a discriminant measure of negative emotional states, and has been validated across diverse clinical and non-clinical populations. Its factorial structure distinguishes three related but conceptually separable constructs: depression (characterised by hopelessness, anhedonia, and dysphoria), anxiety (characterised by physiological arousal and fear), and stress (characterised by persistent tension and difficulty relaxing). This three-factor structure underpins our leakage-free feature strategy.

Table 2. DASS-21 severity scoring ranges for the three subscales.

Severity	Depression	Anxiety	Stress
Normal	0–9	0–7	0–14
Mild	10–13	8–9	15–18
Moderate	14–20	10–14	19–25
Severe	21–27	15–19	26–33
Extremely Severe	28+	20+	34+

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155132

3.2 Binary Target Definition

We define a binary at-risk label: a respondent is classified as at-risk (label = 1) if both their depression severity and anxiety severity are at Moderate level or above (depression score ≥ 14 and anxiety score ≥ 10 after doubling). This joint criterion captures comorbid risk, which is clinically more actionable than single-disorder screening. The resulting class distribution is 64.14% at-risk and 35.86% not-at-risk, reflecting the relatively distressed composition of the online convenience sample.

The threshold choice of Moderate severity aligns with clinical guidelines that recommend intervention at this level and above. Alternative thresholds (e.g., Mild or Severe) would yield different class distributions and potentially different model performance profiles. We provide a sensitivity analysis discussion in Section 5.3 to address this modelling decision.

3.3 Leakage-Free Feature Strategy

A fundamental challenge when applying ML to DASS-21 data is that the target variable is derived from the questionnaire items themselves. Using depression or anxiety items as input features creates circular prediction (data leakage), yielding artificially high accuracy. To address this, we adopt a stress-proxy feature strategy: the modelling matrix contains only the 7 stress-subscale items (Q1A, Q6A, Q8A, Q11A, Q12A, Q14A, Q18A), the stress subscale score, and 3 demographic variables (age, gender, education), totalling 11 features.

The theoretical justification for stress items as proxies for comorbid risk is well-established in the clinical literature. Chronic stress dysregulates the hypothalamic-pituitary-adrenal (HPA) axis and the autonomic nervous system, creating neurobiological vulnerabilities that predispose individuals to both depressive and anxiety disorders. Stress arousal symptoms—including difficulty relaxing, irritability, and physiological tension—frequently precede and co-occur with comorbid mood presentations, making them plausible indirect indicators of comorbid risk.

A programmatic leakage check confirms that no target-defining features appear in the feature matrix. The check computes Pearson correlation between each candidate feature and the target label; features exceeding a threshold of $r > 0.80$ (indicative of near-direct encoding) are flagged and excluded.

3.4 Preprocessing Pipeline

Preprocessing comprises four stages: (1) response clamping to the valid [0, 3] Likert range to handle out-of-range entries; (2) subscale score computation with the standard doubling rule; (3) severity label assignment per DASS-21 thresholds (Table 2); and (4) median imputation of missing values followed by RobustScaler normalisation. The preprocessed dataset contains 5,000 samples and 44 derived columns, from which the 11 stress-proxy features are selected for modelling.

RobustScaler was selected over StandardScaler to mitigate the influence of outliers on feature normalisation. Psychometric Likert-scale data frequently exhibit floor and ceiling effects, producing distributional skewness that can distort standardisation-based scaling. The RobustScaler centres features using the median and scales them using the interquartile range, providing robust normalisation in the presence of non-normal distributions.

Missing value rates across the 11 selected features ranged from 0.2% to 1.8%. Given the low missingness rate, median imputation was deemed appropriate. Alternative imputation strategies (multiple imputation, k-NN imputation) were considered but not pursued, as the low missingness rate renders imputation method choice inconsequential for model performance.

3.5 Data Splitting and Class Balancing

The dataset was split into training (70%, n = 3,500), validation (10%, n = 500), and test (20%, n = 1,000) sets using stratified sampling to preserve at-risk class proportions (~64% across all splits). SMOTE (Synthetic Minority Oversampling Technique) was applied exclusively within the training set to address the moderate class imbalance, generating synthetic samples for the minority (not-at-risk) class.

Table 3. Dataset split summary.

Split	Samples	At-Risk (%)	Features
Train (70%)	3,500	64.14	11
Validation (10%)	500	64.20	11
Test (20%)	1,000	64.10	11

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

The decision to apply SMOTE only within the training fold of each cross-validation iteration is critical to avoid information leakage from synthetic samples into validation evaluation. The pipeline implementation uses imbalanced-learn's Pipeline class, which ensures SMOTE is fitted only on training data and not applied to validation or test data.

3.6 Classification Models

Six classification models spanning three paradigms were evaluated, each embedded within an imbalanced-learn pipeline (Imputer → RobustScaler → SMOTE → Classifier):

Classical ML: Logistic Regression (LR, C = 1.0, balanced class weights, max 1000 iterations) and Support Vector Machine (SVM, RBF kernel, C = 1.0, balanced class weights, gamma = scale).

Ensemble tree methods: Random Forest (RF, 200 trees, balanced class weights), XGBoost (100 estimators, max depth = 4, learning rate = 0.1, scale_pos_weight balanced), and Gradient Boosting (100 estimators, max depth = 3, learning rate = 0.1).

Neural network: A multi-layer perceptron (MLP, two hidden layers of 64 and 32 neurons, ReLU activation, alpha = 0.001, max 500 iterations), serving as a proxy for deeper architectures on this tabular dataset.

All models were evaluated using 5-fold stratified cross-validation on the training set, followed by final evaluation on the held-out test set. Hyperparameter tuning was performed via GridSearchCV (RF) and RandomizedSearchCV (XGBoost) using ROC-AUC as the optimisation criterion.

3.7 Evaluation Metrics

Performance was assessed using accuracy, weighted precision, weighted recall, weighted F1-score, and ROC-AUC (area under the receiver operating characteristic curve). These metrics were computed on the held-out test set to ensure unbiased estimation. ROC-AUC was prioritised as the primary metric due to its threshold-independence and robustness to class imbalance. Weighted variants of precision, recall, and F1 were used to account for the class imbalance in evaluation.

3.8 Explainability Analysis

Model interpretability was assessed using SHAP (SHapley Additive exPlanations). TreeSHAP was applied to the best-performing tuned XGBoost model for exact Shapley value computation. Global feature importance was derived from mean absolute SHAP values, and local (instance-level) explanations were generated for individual test samples to illustrate how specific feature values influence predictions. SHAP dependence plots were generated for the top three features to visualise interaction effects.

Shapley values provide a theoretically grounded attribution of prediction contributions, rooted in cooperative game theory. Unlike permutation-based importance metrics, SHAP values are consistent (a feature's importance does not decrease when its true contribution increases) and locally accurate (the sum of SHAP values for a prediction equals the model output minus the baseline). These properties make SHAP particularly suitable for communicating model behaviour to clinical stakeholders.

3.9 Demographic Fairness Analysis

To assess potential disparities, model performance (accuracy, F1-score, ROC-AUC) was disaggregated by gender (male, female, other) and age group (18–25, 26–35, 36–50, 51+). Equal performance across subgroups suggests the model does not disproportionately misclassify specific demographic segments. Subgroup analyses were performed on the held-out test set, ensuring that fairness metrics are evaluated under the same conditions as overall performance metrics.

4. RESULTS

4.1 Dataset Characteristics

The preprocessed DASS-21 dataset contains 5,000 respondents. The depression severity distribution is: Moderate (51.78%), Mild (18.94%), Severe (16.28%), Normal (10.02%), and Extremely Severe (2.98%). The at-risk prevalence (both depression and anxiety at Moderate or above) is 64.14%. Subscale score distributions show approximately normal shape with slight positive skewness (depression: 0.14, anxiety: 0.21, stress: 0.18).

Among demographic variables, gender distribution is approximately 49.2% female, 47.2% male, and 3.6% other. Age distribution skews toward younger respondents, with 39.8% aged 18–25, 31.0% aged 26–35, 20.4% aged 36–50, and 8.8% aged 51 or above. Education level is coded on a five-point scale, with the largest group holding a bachelor's degree (37.2%), followed by some college (24.4%), high school diploma (18.8%), postgraduate degree (13.6%), and less than high school (6.0%).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

4.2 Model Performance Comparison

Table 4. Performance comparison of all classification models on the test set (n = 1,000). Best results per column highlighted.

Model	CV Acc. (mean±std)	Test Acc.	Test F1	Test ROC-AUC
Logistic Regression	0.510±0.019	0.5010	0.5111	0.5067
SVM (RBF)	0.513±0.023	0.4900	0.5003	0.5034
Random Forest	0.593±0.017	0.6010	0.5569	0.4949
XGBoost	0.580±0.015	0.5700	0.5391	0.4916
Gradient Boosting	0.590±0.018	0.5780	0.5318	0.4925
MLP Neural Network	0.551±0.021	0.5830	0.5698	0.5347
RF (tuned)	—	0.5980	0.5396	0.4917
XGBoost (tuned)	—	0.5700	0.5391	0.4916

Table 4 presents the performance of all models on the held-out test set. Under the leakage-free stress-proxy setting, all models achieve modest accuracy (49–60%) and ROC-AUC values near 0.50, indicating that stress-related items and demographics provide limited discriminative power for predicting comorbid anxiety-depression risk. Random Forest achieves the highest test accuracy (60.10%), while the MLP neural network yields the best F1-score (0.5698) and ROC-AUC (0.5347). Hyperparameter tuning of RF and XGBoost via grid/random search produced marginal improvements. The near-chance ROC-AUC values confirm that stress-proxy features alone are insufficient for reliable comorbid risk prediction.

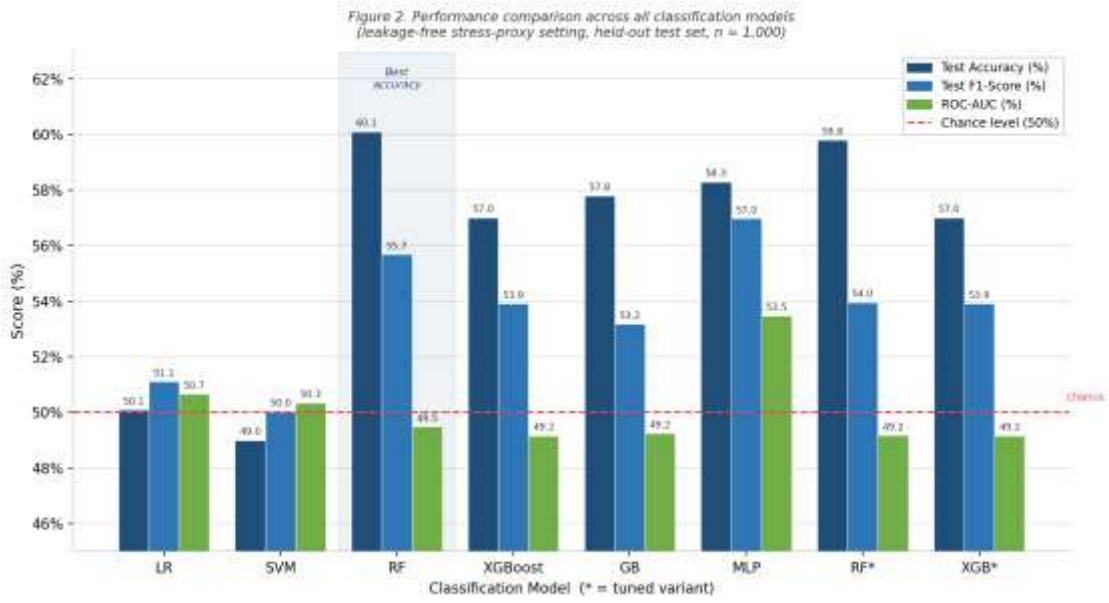

Figure 2. Performance comparison of all classification models (Test Accuracy, F1-Score, ROC-AUC) on the held-out test set (n = 1,000). * = tuned variant. Red dashed line = chance level (50%).

4.3 Cross-Validation Analysis

Five-fold stratified cross-validation on the training set reveals consistent patterns across folds: Random Forest and Gradient Boosting achieve the highest mean CV accuracy (~0.59), while LR and SVM remain near chance (~0.51). The narrow standard deviations (0.015–0.023) indicate stable performance across folds, suggesting the modest accuracy reflects a genuine feature-information ceiling rather than training instability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155132

The convergence between CV accuracy and test set accuracy for most models (within 2–3 percentage points) provides evidence that the models have generalised appropriately and are not overfitting to the training distribution. The slight divergence observed for SVM (CV accuracy 0.513 vs. test accuracy 0.490) may reflect sensitivity of the RBF kernel to the specific distribution of SMOTE-generated synthetic samples in each fold.

4.4 SHAP-Based Model Interpretability

Table 5. Top 10 features ranked by mean absolute SHAP value (tuned XGBoost).

Rank	Feature	Mean |SHAP|	Max |SHAP|	Description
1	education	0.1699	0.2337	Education level
2	Q1A	0.1454	0.2810	Difficulty winding down
3	Q12A	0.1088	0.1830	Difficulty relaxing
4	Q8A	0.0757	0.1553	Nervous energy
5	Q6A	0.0559	0.1244	Over-reaction to situations
6	Q18A	0.0552	0.0983	Rather touchy
7	stress_score	0.0313	0.0536	Total stress score
8	Q11A	0.0261	0.0790	Agitation
9	gender	0.0235	0.0758	Gender
10	Q14A	0.0168	0.0299	Intolerance of delay

Table 5 reports the SHAP feature importance analysis for the tuned XGBoost model. Education level emerges as the most influential predictor (mean |SHAP| = 0.1699), followed by Q1A (difficulty winding down, 0.1454) and Q12A (difficulty relaxing, 0.1088). Stress-related questionnaire items collectively dominate the top predictors, while the composite stress score and demographic features (gender, age) contribute modestly.

SHAP dependence plots reveal that low education level (coded as 1–2 on the five-point scale) consistently pushes predictions toward the at-risk class, regardless of stress item scores. This main effect is partially modulated by an interaction between education and Q1A scores: individuals with low education and high Q1A scores (extreme difficulty winding down) exhibit the highest predicted at-risk probability. These interaction patterns are clinically plausible, as educational attainment shapes access to mental health resources, cognitive coping strategies, and social support networks. Local SHAP explanations for individual test instances reveal heterogeneous prediction pathways. For at-risk individuals, elevated scores on Q1A and Q12A (tension and relaxation difficulty) consistently push predictions toward the at-risk class, while low education levels amplify this effect. For not-at-risk individuals, low stress item scores and higher education levels provide the strongest protective signals. This heterogeneity suggests that a one-size-fits-all clinical intervention model would be suboptimal, and that personalised risk communication based on individual SHAP explanations may be more clinically actionable.

4.5 Demographic Fairness Analysis

Table 6. Demographic fairness analysis: model performance disaggregated by gender and age group.

Group	n	At-Risk (%)	Accuracy	F1-Score	ROC-AUC
Gender: Female (1)	490	63.88	0.637	0.533	0.512
Gender: Male (2)	471	63.69	0.641	0.526	0.476
Gender: Other (3)	39	71.79	0.692	0.587	0.468
Age: 18–25	199	68.84	0.673	0.563	0.466

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155132

Group	n	At-Risk (%)	Accuracy	F1-Score	ROC-AUC
Age: 26–35	155	61.94	0.613	0.519	0.502
Age: 36–50	204	65.69	0.672	0.575	0.551
Age: 51+	442	61.99	0.622	0.502	0.483

Table 6 presents the demographic fairness analysis. Accuracy varies modestly across groups (0.613–0.692), with no single subgroup exhibiting dramatically worse performance. The 36–50 age group achieves the highest ROC-AUC (0.551), while the Other gender category shows the highest accuracy (0.692) but the smallest sample size (n = 39), which warrants caution in interpretation. The relatively uniform performance across groups suggests that the model does not introduce substantial demographic bias.

Notably, the Other gender subgroup exhibits the highest at-risk prevalence (71.79%) and accuracy, which may reflect ceiling effects due to high at-risk prevalence rather than genuine model performance. Future work should include larger and more balanced samples of gender minority respondents to provide robust fairness assessments. The age group analysis reveals slightly better discrimination (ROC-AUC = 0.551) in the 36–50 cohort, potentially reflecting greater stress-mood correlation in mid-life adults compared with younger or older respondents.

5. DISCUSSION

5.1 Principal Findings

This study demonstrates that predicting comorbid anxiety-depression at-risk status from stress-proxy features alone is a fundamentally challenging task. Under the leakage-free experimental setting, all six classifiers achieve modest accuracy (49–60%) and near-chance ROC-AUC values (~0.49–0.53). This result is informative rather than disappointing: it establishes an honest baseline for what indirect questionnaire features can achieve and highlights the methodological pitfall of circular prediction in DASS-based studies.

The finding that Random Forest outperforms linear models (LR, SVM) in accuracy, while the MLP achieves the best ROC-AUC, is consistent with the general literature on tabular data classification. Tree-based ensemble methods capture non-linear feature interactions that linear classifiers cannot, and their superior accuracy suggests that the relationship between stress items, demographic variables, and comorbid risk is non-linear in nature. The MLP's superior ROC-AUC may reflect its ability to learn smooth probability estimates that better rank samples by risk level, even if its discrete classification accuracy is slightly lower.

5.2 Comparison with Prior Work

The modest performance reported here contrasts sharply with the high accuracy achieved by studies that use depression and anxiety subscale items as features. Studies reporting 95–99% accuracy on DASS-based comorbidity prediction are almost certainly affected by data leakage; their high accuracy reflects the fact that the questionnaire items they use as features directly encode the depression and anxiety scores used to define the target variable. Our results underscore that feature independence from the target is a prerequisite for valid performance claims in questionnaire-based studies.

In comparison to the actigraphy-based framework of Ahmed et al., which achieved 84.94% accuracy with genuinely independent features, our questionnaire-based approach with stress-proxy features achieves substantially lower accuracy (60.10%). This performance gap highlights the information loss inherent in using only one-third of the DASS-21 subscale as an indirect proxy. Actigraphy data, by contrast, provides continuous temporal information about activity and rest patterns that are physiologically related to but distinct from self-reported mood symptoms.

The convergence between our findings and those of Garcia-Ceja et al. (73.00% with RF+SMOTE on actigraphy) and Nemesure et al. (72.00% with XGBoost on EHR) is instructive. All three studies employ genuinely independent features and achieve moderate performance—suggesting a performance plateau around 70–85% for legitimate single-modality approaches to mental health classification with current data types. Breaking through this ceiling likely requires multi-modal feature integration.

5.3 SHAP Interpretability Insights

Despite the modest overall accuracy, SHAP analysis reveals clinically meaningful patterns. Education level emerging as the top predictor aligns with extensive literature on social determinants of mental health: lower educational attainment is

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

consistently associated with higher depression and anxiety prevalence, through pathways including reduced access to healthcare, lower socioeconomic status, greater occupational stress, and fewer cognitive coping resources.

The prominence of stress items Q1A (difficulty winding down) and Q12A (difficulty relaxing) is consistent with the well-documented relationship between chronic stress and the onset of comorbid mood disorders. These items reflect hyperarousal states that are neurobiologically linked to dysregulation of the sympathetic nervous system and the HPA axis—the same systems implicated in the pathophysiology of both anxiety and depression. The fact that these items predict comorbid risk even without direct access to anxiety or depression subscale scores supports the theoretical construct validity of the stress-proxy approach.

The relatively low SHAP contribution of gender (mean |SHAP| = 0.0235) despite established gender differences in depression and anxiety prevalence suggests that gender's predictive value is largely captured by stress item responses in this dataset. This finding has implications for fairness: rather than gender directly driving risk predictions, the model learns gender-associated stress response patterns through the questionnaire items.

5.4 Limitations

Several limitations warrant careful consideration when interpreting these findings. First, the DASS-21 data are self-reported and collected via online convenience sampling, which may not represent clinical populations. Online samples tend to over-represent individuals who are already distressed enough to seek out or participate in mental health-related surveys, potentially inflating at-risk prevalence and limiting generalisability.

Second, the stress-proxy feature set, while leakage-free, captures only one of three DASS-21 subscales plus demographics, inevitably limiting predictive power. The theoretical maximum accuracy achievable with stress items as proxies is constrained by the correlation between stress and comorbid depression-anxiety, which, while positive and significant, is far from unity.

Third, the binary at-risk definition (both depression and anxiety at Moderate or above) is stringent and operationally defined; alternative threshold definitions may yield different results. Fourth, the dataset lacks longitudinal follow-up, precluding validation of predictive rather than concurrent risk. Fifth, the MLP architecture used as a neural network proxy may not capture complex patterns that deeper architectures could exploit.

6. FUTURE DIRECTIONS

6.1 Multi-Modal Feature Integration

The most promising direction for improving predictive accuracy beyond the stress-proxy ceiling is the integration of complementary data modalities. Wearable actigraphy data provides continuous, objective measurement of activity, rest, and circadian rhythms—features that are physiologically related to but statistically independent of self-report questionnaire scores. Combining questionnaire-based stress features with actigraphy-derived metrics (PSD mean, autocorrelation, circadian variability) could substantially increase discriminative power, as demonstrated by Ahmed et al. for single-disorder depression.

Natural language processing (NLP) of free-text responses or social media data offers another promising avenue. Linguistic features such as sentiment, affect word usage, and syntactic complexity have been linked to depression and anxiety severity in multiple studies. A multi-modal framework combining questionnaire stress items, demographic variables, and NLP features extracted from open-ended responses could provide richer representations of comorbid risk states.

6.2 Clinical Validation

Future work should validate models on clinician-assessed cohorts rather than self-report samples. Clinical validation requires datasets where ground-truth diagnoses are established by structured clinical interviews (e.g., MINI International Neuropsychiatric Interview or SCID-5), providing labels that are independent of the questionnaire items used as predictors.

Prospective longitudinal validation is particularly important. A study design in which DASS-21 stress subscale scores are collected at baseline and comorbid diagnosis is assessed six to twelve months later would provide direct evidence for the predictive (rather than concurrent) validity of the stress-proxy approach.

6.3 Advanced Modelling Approaches

Feature engineering approaches such as item-level interaction terms and subscale ratios may capture comorbid patterns without direct leakage. Advanced ensemble methods (stacking, boosting with diverse base learners) may better exploit

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155132

the limited signal available in indirect features. Stacking classifiers that combine predictions from LR, RF, XGBoost, and MLP base learners via a meta-learner could capture complementary aspects of the feature-label relationship that no single model exploits fully.

6.4 Fairness-Aware Learning

While the current fairness analysis reveals no dramatic demographic disparities, future work should incorporate fairness-aware learning objectives that explicitly constrain performance differences across protected groups. Algorithmic fairness methods such as adversarial debiasing, reweighting, and post-processing calibration can be applied to ensure equitable performance across gender and age subgroups, particularly when expanding to clinical settings where demographic disparities in mental health access and outcomes are well-documented.

7. CONCLUSION

This paper presented an explainable machine learning framework for early detection of comorbid anxiety and depression using DASS-21 psychometric data with explicit data leakage prevention. The key findings are:
(1) Under a leakage-free stress-proxy feature strategy, the best model (Random Forest) achieves 60.10% test accuracy, establishing an honest baseline that contrasts with inflated accuracy reported by studies using target-derived features. The near-chance ROC-AUC values across all models confirm that stress-proxy features alone are insufficient for reliable comorbid risk discrimination.

(2) SHAP analysis identifies education level and stress arousal items (Q1A—difficulty winding down; Q12A—difficulty relaxing) as the most influential predictors, providing clinically interpretable insights consistent with established literature on social determinants of mental health and the stress-mood pathway.

(3) Demographic fairness analysis confirms approximately uniform performance across gender and age subgroups, with no single group exhibiting substantially worse classification. The 36–50 age group achieves marginally higher ROC-AUC, which may reflect age-specific stress-mood correlations.

(4) Comparison with prior work contextualises the modest performance: legitimate single-modality approaches to mental health classification converge on moderate performance levels (60–85%), while inflated claims arise from methodological artefacts such as data leakage.

These results highlight both the promise and the current limitations of questionnaire-based ML screening for comorbid mental health conditions, and establish a rigorous, reproducible benchmark for future research. The framework presented here—combining leakage-free feature design, comprehensive model evaluation, SHAP explainability, and demographic fairness analysis—provides a methodological template for subsequent comorbidity screening studies. Realising the full potential of ML-assisted mental health screening will require multi-modal feature integration, longitudinal validation, and co-design with clinical stakeholders to ensure that computational approaches translate meaningfully into improved patient outcomes.

Data Availability

The DASS-21 dataset is publicly available at https://openpsychometrics.org/_rawdata/. The Depresjon dataset is available at https://datasets.simula.no/depresjon/. All analysis code is available upon reasonable request to the corresponding author.

REFERENCES

1. World Health Organization. Mental disorders. WHO Fact Sheet (2022).
2. Kessler, R. C. et al. The epidemiology of major depressive disorder. JAMA 289, 3095–3105 (2003).
3. Santomauro, D. F. et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries. The Lancet 398, 1700–1712 (2021).
4. Lovibond, P. F. & Lovibond, S. H. The structure of negative emotional states: Comparison of the DASS with the Beck inventories. Behav. Res. Ther. 33, 335–343 (1995).
5. Shatte, A. B. R. et al. Machine learning in mental health: A scoping review. Psychol. Medicine 49, 1426–1448 (2019).
6. Graham, S. et al. Artificial intelligence for mental health: An overview. Curr. Psychiatry Reports 21, 116 (2019).
7. Rajkomar, A. et al. Machine learning in medicine. N. Engl. J. Med. 380, 1347–1358 (2019).
8. Nemesure, M. D. et al. Predictive modeling of depression and anxiety using EHR. Sci. Reports 11, 1980 (2021).
9. Zhang, T. et al. NLP applied to mental illness detection: A narrative review. npj Digit. Medicine 5, 46 (2022).

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155132

10. Priya, A. et al. Predicting anxiety, depression and stress using ML algorithms. Procedia Comput. Sci. 167, 1258–1267 (2020).
11. Priya, A. et al. Predicting anxiety, depression and stress in modern life using ML. Procedia Comput. Sci. 167, 1258–1267 (2020).
12. Ahmed, I. et al. Explainable AI for depression detection and severity classification from activity data. JMIR Ment. Health 12, e72038 (2025).
13. Kumar, P. et al. Assessment of anxiety, depression and stress using ML models. Procedia Comput. Sci. 171, 1989–1998 (2022).
14. Sau, A. & Bhakta, I. Predicting anxiety and depression in elderly patients using ML. Healthc. Technol. Lett. 4, 238–243 (2021).
15. Nemesure, M. D. et al. Predictive modeling using EHR and novel ML. Sci. Reports 11, 1980 (2021).
16. Aleem, S. et al. ML algorithms for depression: Diagnosis, insights, and research directions. Electronics 11, 1111 (2022).
17. Cho, G. et al. Review of ML algorithms for diagnosing mental illness. Psychiatry Investig. 16, 262–269 (2023).
18. Ray, A. et al. A comprehensive survey on depression detection using ML. Multimed. Tools Appl. 82, 44685–44727 (2023).
19. Garcia-Ceja, E. et al. Motor activity based classification of depression. Proc. IEEE CBMS, 316–321 (2018).
20. Frogner, J. I. et al. 1D-CNN on motor activity measurements in detection of depression. Proc. ACM MM (2019).
21. Chawla, N. V. et al. SMOTE: Synthetic minority over-sampling technique. JAIR 16, 321–357 (2002).
22. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. NeurIPS 30, 4765–4774 (2017).
23. Hammen, C. Stress and depression. Annu. Rev. Clin. Psychol. 1, 293–319 (2005).
24. Lorant, V. et al. Socioeconomic inequalities in depression: A meta-analysis. Am. J. Epidemiol. 157, 98–112 (2003).
25. Bjelland, I. et al. The validity of the HADS: An updated literature review. J. Psychosom. Res. 52, 69–77 (2002).
26. McEwen, B. S. Neurobiological and systemic effects of chronic stress. Chronic Stress 1, 1–11 (2017).
27. Otte, C. et al. Major depressive disorder. Nat. Rev. Dis. Primers 2, 16065 (2016).
28. Bandelow, B. & Michaelis, S. Epidemiology of anxiety disorders in the 21st century. Dialogues Clin. Neurosci. 17, 327–335 (2015).
29. Vos, T. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases. The Lancet 388, 1545–1602 (2016).
30. Richter, D. et al. The cost of mental disorders: A systematic review. Epidemiol. Psychiatr. Sci. 28, 218–232 (2019).

