An Explainable Machine Learning Framework for Early Mental Health Risk Detection among University Students in Fragile and Conflict-Affected Settings

[1] Mr. Abdelatti Blg [2] Mr. Muner Athaba [3] Mr. Ahmed Ashlam

https://orcid.org/ 0009-0008-5404-9527 1 Author
https://orcid.org/ 0009-0006-9688-8156 2 Author
https://orcid.org/ 0009-0002-4140-8131 3 Author

[1 2] Department of Computer Science, Faculty of Science, University of Zawia (ZU), Libya
[3] Ministry of Technical and Vocational Education, Libya
[1] a.blg@zu.edu.ly [2] m.athaba@zu.edu.ly [3] ahmed_nea@yahoo.com

تطوير نموذج للتعلم الآلي القابل للتفسير للكشف المبكر عن مخاطر الصحة النفسية بين طلاب الجامعات في البيئات الهشة والمتأثرة بالنزاعات

أ. عبدالعاطي بلق [1] أ. منير الضبع [2] أ. أحمد ألاشلم [3]
قسم علوم الحاسوب، كلية العلوم، جامعة الزاوية، ليبيا [1 2]
وزارة التعليم التقني، ليبيا [3]

تاريخ الاستلام: 2026-06-01، تاريخ القبول: 2026-06-10، تاريخ النشر: 2026-06-25

Abstract:

Mental health issues among university students continue to receive growing attention worldwide. These challenges are often more pronounced in fragile and conflict-affected settings (FCAS), where instability, displacement, and socioeconomic pressures can increase psychological vulnerability. In such environments, the early identification of students who may be at elevated mental health risk is important for enabling timely and appropriate support. Although machine-learning techniques have shown considerable potential in mental health analytics, many existing studies place greater emphasis on predictive performance than on model transparency and practical usability. As a result, the adoption of these models in real-world educational settings remains challenging, particularly when decision-makers require explanations that are easy to understand and trust. This study develops an explainable machine-learning framework for early mental health risk stratification among university students in FCAS contexts. The proposed framework combines demographic, academic, socioeconomic, psychological, and contextual indicators within a structured modelling pipeline designed to support transparency, calibration, and practical decision-making. Several classification algorithms were evaluated using stratified five-fold cross-validation. Among the evaluated models, logistic regression provided the most favourable balance between predictive performance and interpretability. The final model achieved an accuracy of 0.935, a macro-averaged F1 score of 0.862, and a one-versus-rest AUC of 0.997 for identifying high-risk students. Calibration analysis further indicated reliable probability estimates, with a Brier score of 0.024, while decision curve analysis demonstrated a consistent net benefit across a range of clinically relevant decision thresholds. To better understand the model's behaviour, SHAP-based explainability analysis was performed. The results showed that psychological indicators, particularly stress, depression, and anxiety, were the strongest contributors to high-risk predictions, whereas demographic variables had a comparatively limited influence. Overall, the findings suggest that lightweight and interpretable machine-learning models can provide effective decision support for early mental health screening, offering a practical solution for resource-constrained university environments.

Keywords:
Mental health risk detection, explainable machine learning, SHAP, early warning systems, university students, fragile and conflict-affected settings, decision curve analysis

الملخص:

تمثل تحديات الصحة النفسية بين طلاب الجامعات مصدر قلق عالمي متزايد، لا سيما في البيئات الهشة والمتأثرة بالنزاعات(FCAS) ، حيث يؤدي التعرض لعدم الاستقرار والنزوح والضغوط الاجتماعية والاقتصادية إلى زيادة الهشاشة النفسية. لذلك، يُعدّ الكشف المبكر عن الطلاب الأكثر عرضة لمخاطر مشكلات الصحة النفسية أمرًا ضروريًا لتقديم الدعم المناسب وفي الوقت المناسب.

وعلى الرغم من أن أساليب التعلم الآلي أظهرت إمكانات واعدة في تحليل قضايا الصحة النفسية، فإن العديد من النماذج الحالية تركز على الدقة التنبؤية على حساب قابلية التفسير وإمكانية التطبيق العملي.

تقترح هذه الدراسة إطارًا قائمًا على التعلم الآلي القابل للتفسير للتصنيف المبكر لمستويات مخاطر الصحة النفسية بين طلاب الجامعات في البيئات الهشة والمتأثرة بالنزاعات. يدمج الإطار مجموعة من المؤشرات الديموغرافية والأكاديمية والاجتماعية والاقتصادية والنفسية والسياقية ضمن خط أنابيب نمذجة منظم يركز على الشفافية، ومعايرة التنبؤات، وفائدة اتخاذ القرار. وقد تم تقييم عدة مصنفات باستخدام أسلوب التحقق المتقاطع الطبقي خماسي الطيات(Stratified Five-Fold Cross-Validation) ، حيث برز نموذج الانحدار اللوجستي باعتباره النموذج الأكثر فعالية وقابلية للتفسير.

حقق النموذج المختار دقة بلغت **0.935**، ودرجة **F1**كلية (Macro-Averaged) بلغت **0.862**، ومساحة تحت المنحنى (**AUC**) بطريقة الواحد مقابل البقية (One-Versus-Rest) بلغت **0.997** للكشف عن الفئة عالية الخطورة. كما أظهر تحليل المعايرة موثوقية عالية في تقديرات الاحتمالات (درجة **Brier = 0.024**)، في حين أكد تحليل منحنى القرار (**Decision Curve Analysis**) وجود منفعة صافية متسقة عبر نطاق واسع من العتبات ذات الأهمية السريرية.

ولدعم قابلية التفسير، تم إجراء تحليلات قائمة على تقنية **SHAP**لتحديد العوامل الرئيسة المؤثرة في التنبؤات المتعلقة بالفئة عالية الخطورة. وأظهرت النتائج أن المؤشرات النفسية، وبشكل خاص **التوتر والاكتئاب والقلق**، تهيمن على عملية اتخاذ القرار في النموذج، بينما تسهم الخصائص الديموغرافية بدرجة محدودة جدًا. وبصورة عامة، تؤكد النتائج أن النماذج الخفيفة والقابلة للتفسير قادرة على تحقيق أداء تنبؤي

قوي وتوفير دعم فعّال لاتخاذ القرار، مما يوفر أساسًا عمليًا للكشف المبكر عن مخاطر الصحة النفسية في البيئات الجامعية محدودة الموارد.

الكلمات المفتاحية: الكشف عن مخاطر الصحة النفسية، التعلم الآلي القابل للتفسير، SHAP، أنظمة الإنذار المبكر، طلاب الجامعات، البيئات الهشة والمتأثرة بالنزاعات، تحليل منحنى القرار.

1. Introduction

In many universities, student mental health is no longer a secondary issue. During our work with university students, we observed that academic pressure is only one part of the problem. Some students also face financial difficulty, family obligations, unstable social conditions, and uncertainty about their future. These issues may gradually affect their wellbeing and increase symptoms related to stress, anxiety, or depression. Similar concerns have been reported in recent studies involving university students from different contexts (Hasan et al., 2025; Yu et al., 2025). The problem is more visible in fragile and conflict-affected settings, where normal study life can be interrupted by instability, displacement, and wider social disruption.

Early recognition of these problems is therefore important. Universities often become aware of students' difficulties only after the situation has already become serious. Machine learning may help in this area by identifying patterns that are difficult to notice manually. Previous studies have used supervised learning models to predict depression, anxiety, and stress from psychological and behavioural survey data (Biswas et al., 2025; Opar et al., 2024; Baba and Bunji, 2023). However, most of these studies were conducted in relatively stable educational environments. Their findings may not fully reflect the situation of students living in fragile or crisis-affected contexts. Recent studies have also used hybrid and ensemble models, including DA-TabSVM and gradient-boosting methods, to improve classification performance (Arora and Dahiya, 2025; Zhai et al., 2025). These models are useful from a technical point of view, especially when they achieve high accuracy. However, in student wellbeing, a correct prediction is not the only requirement. The university also needs to understand the reason behind the prediction before taking any support action. For this reason, explainability is important, especially when the system is used by student services, academic advisors, or university administrators rather than clinical specialists (Drira et al., 2024; Campbell et al., 2024). The current literature still gives limited attention to fragile and conflict-affected settings. Most studies focus on individual psychological measures, while the surrounding context is often treated as background information or ignored completely. In practice, factors such as instability, displacement, and socioeconomic hardship may shape students' mental health in important ways. Another limitation is that many studies report accuracy and other performance metrics but provide less discussion about model explanation, ethical use, or practical deployment in universities with limited resources (Nath et al., 2025 and El Morr et al., 2024).

Based on this gap, this paper develops an explainable machine-learning framework for early mental health risk detection among university students in fragile and conflict-affected settings. The framework uses psychological, socio-academic, and contextual indicators together. It is not intended to diagnose mental illness. Rather, it is designed to support early risk stratification and help universities identify students who may need timely support.

2. Related Work

Research on student mental health prediction has expanded considerably in recent years. Many studies have explored the use of machine-learning techniques to identify symptoms of depression, anxiety, and stress among university students. Most of these studies rely on survey-based data that combine psychological, demographic, and behavioural information. For example, Biswas et al. (2025) and Hasan et al. (2025) applied classification models using demographic and self-reported indicators, while Yu et al. (2025) and Opar et al. (2024) reported that behavioural and academic variables can also contribute to the identification of students who may be experiencing depressive symptoms.

Researchers have also explored more advanced modelling approaches. Arora and Dahiya (2025) proposed the DA-TabSVM model for depression and anxiety detection, demonstrating improvements over several baseline methods. Similarly, Chai et al. (2025) demonstrated that machine learning predictions can support targeted prevention strategies by helping institutions prioritise intervention resources based on estimated risk levels. Despite these advances, most current work focuses on improving predictive performance, with less attention paid to the practical application of these models in educational settings. Another area of increasing interest is interpretability. Campbell et al. (2024) showed that interpretable AI techniques can improve the usability and reliability of mental health prediction systems in university settings. Drira et al. (2024) further emphasised the importance of transparency and interpretability when machine-learning models are used to support decisions related to student wellbeing. This concern is consistent with broader findings in educational machine-learning research, where explainability, responsible deployment, and practical applicability have increasingly been recognised as important considerations alongside predictive performance (Blg et al, 2026). Although explainable approaches have been investigated in clinical applications, including self-harm detection using language models (Martínez-Romo et al., 2025), their adoption within higher education remains relatively limited. The context in which prediction models are developed also deserves attention. Most published studies are conducted in stable educational environments, with limited consideration of fragile and conflict-affected settings. In practice, students in such environments may face displacement, financial hardship, social instability, and disruptions to their educational activities. These contextual factors can affect mental health, but they are often overlooked in predictive models. Al-Dakka (2025) noted that research on mental health using artificial intelligence in conflict-affected areas remains relatively limited. While recent studies have examined digital mental health interventions and risk-level prediction in broader populations (Suh and Yoo, 2025) and (El Morr et al., 2024), the integration of contextual vulnerability with student-level indicators remains underexplored.

Building on these observations, the present study focuses on early mental health risk detection among university students in fragile and conflict-affected settings. In addition to predictive performance, the proposed framework places particular emphasis on interpretability, contextual awareness, and the responsible use of machine-learning models in higher education.

3. Materials and Datasets

3.1. Study Design and Data Overview

This work follows a quantitative research design based on supervised machine learning. The main purpose is to develop and evaluate an explainable model that can support early identification of mental health risk among university students in fragile and conflict-affected settings. The study is not intended to provide clinical diagnosis or estimate disease prevalence. Instead, it is designed as a methodological and decision-support study that can help universities recognise students who may need early support. In this setting, mental health risk is treated as a multi-class classification problem. Students are grouped into low, moderate, or high-risk categories based on the available indicators. This approach is consistent with previous studies on student mental health prediction, where machine-learning models were used to support early risk stratification rather than clinical decision-making (Biswas et al., 2025; Hasan et al., 2025; Baba and Bunji, 2023).

The dataset representation was guided by two sources. First, it follows the structure used in previous student mental health studies and validated psychological instruments. Second, it extends this structure by including engineered and contextual features that are relevant to fragile and conflict-affected environments. This combination allows the analysis to consider both individual student characteristics and the wider conditions surrounding the student.

3.2 Dataset

The study uses a structured student-level dataset collected from a higher-education context affected by fragility and conflict. The dataset includes demographic, academic, socioeconomic, contextual, and psychological variables. These variables were used to construct a three-level mental health risk label: low, moderate, and high risk. Before analysis, the data were anonymised, and all modelling procedures were conducted using de-identified records. The dataset was organised around two main groups of information. The first group represents student-level indicators, including psychological, demographic, academic, and socioeconomic characteristics. The second group represents contextual stress indicators, which reflect broader forms of instability that may be present in fragile and conflict-affected settings.

Using these two groups together allows the model to capture more than individual psychological vulnerability alone. It also considers wider contextual pressures that may influence students' mental health, while avoiding the use of identifiable personal information.

3.3. Feature Composition

When defining the feature set, attention was given not only to predictive performance but also to interpretability and ethical considerations. This reflects recent recommendations in responsible AI research, which encourage the development of transparent and trustworthy machine-learning systems for student mental health applications (Drira et al., 2024; Campbell et al., 2024).

3.3.1 Student-Level Features

The student-level variables covered four main areas. Demographic information included age and gender, while academic information described the programme of study and the student's academic performance level. Socioeconomic conditions were represented through indicators such as perceived financial stability and living conditions. Psychological risk was captured using measures related to depression and anxiety, derived from validated instruments.

These types of variables are commonly used in student mental health prediction studies because they are closely linked to mental health outcomes in university populations (Hasan et al., 2025; Zhai et al., 2025).

3.3.2 Contextual Features

Contextual features were added to represent the conditions surrounding students in fragile and conflict-affected settings. These variables were used at an aggregate level and reflected broader stressors such as instability, displacement, and disruption to normal life. Although such factors are often missing from student mental health prediction studies, they are important because the surrounding environment can shape students' psychological well-being. Al-Dekah (2025) also identified contextual vulnerability as an underrepresented dimension in current AI-based mental health research.

4. Methodology

4.1 Overall Study Design

This study adopts a supervised machine-learning approach to support early mental health risk stratification among university students in fragile and conflict-affected settings (FCAS). The task is formulated as a three-class classification problem, distinguishing between low, moderate, and high risk levels. Importantly, the framework is intended as an early-warning and decision-support tool, rather than a diagnostic system, with the aim of enabling timely and proportionate institutional interventions. An overview of the end-to-end modelling pipeline is presented in Figure 1, which illustrates the sequence of preprocessing, feature handling, class-imbalance mitigation, model training, evaluation, and explainability.

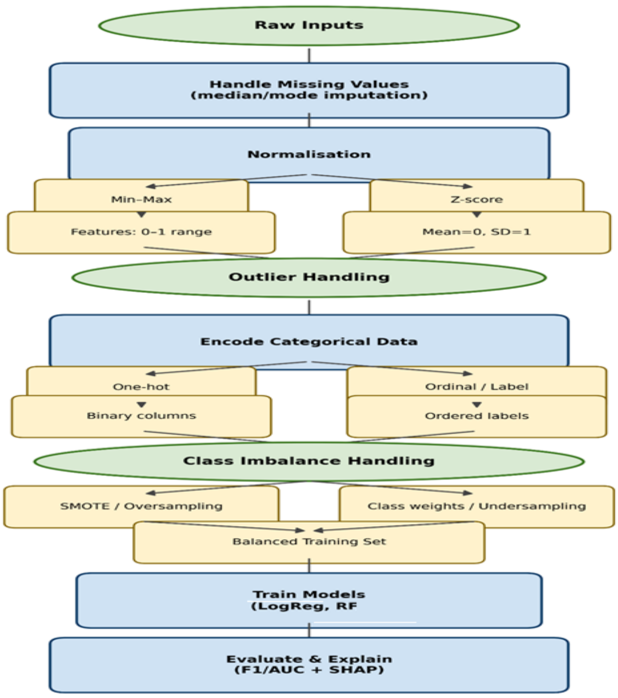

Figure 1. Overview of the proposed explainable machine-learning pipeline, including data preprocessing, class imbalance handling, model training, evaluation, and SHAP-based explainability.

4.2 Data Characteristics and Learning Task

The key characteristics of the dataset and modelling setup are summarised in **Table 1**. The learning task involves multi-class mental health risk prediction, using heterogeneous student-level indicators derived from demographic, academic, psychological, socioeconomic, and contextual sources.

Aspect	Description
Data Source	University student survey and contextual indicators collected from a fragile and conflict-affected setting
Learning Task	Multi-class classification of mental health risk (Low / Moderate / High)
Target Class of Interest	High-risk mental health status
Feature Representation	Integrated demographic, academic, socioeconomic, psychological, contextual, and engineered features
Data Format	Tabular structured data with aggregated contextual indicators
Missing Data Handling	Median imputation for numerical variables, mode imputation for categorical variables
Feature Scaling	Min–max normalisation or z-score standardisation, depending on model requirements
Class Imbalance Handling	Class weighting and/or SMOTE applied within training folds
Models Evaluated	Logistic Regression, Random Forest
Cross-Validation Strategy	Stratified five-fold cross-validation
Evaluation Metrics	Accuracy, Macro-averaged F1 score, One-vs-Rest AUC (high-risk class)
Probability Calibration	Calibration curves and Brier score assessment
Clinical Utility Assessment	Decision Curve Analysis for high-risk prediction
Explainability Method	SHAP-based global and instance-level feature attribution
Intended Use	Early mental health risk stratification to support targeted university interventions

Table 1. Summary of data characteristics, modelling assumptions, and evaluation design.

The dataset integrates contextual indicators relevant to FCAS environments, supporting a more realistic representation of student stress exposure. Model performance is evaluated using accuracy, macro-averaged F1-score, and area under the ROC curve (AUC), alongside calibration and decision-analytic metrics to assess practical utility.

4.3 Feature Representation

Table 2 presents the main feature categories used in the study, together with representative variables. Demographic variables, such as age and gender, describe the basic characteristics of the students. Academic variables capture information related to enrolment and academic performance. The psychological variables, including depression, anxiety, and stress measures, represent the main indicators used to assess mental health risk. Socioeconomic and contextual variables were included to reflect pressures that may affect students' well-being, such as financial instability, displacement, and exposure to fragile living conditions. In addition, engineered aggregate variables, such as contextual stress indices, were used to summarise multiple sources of risk in a form that remains interpretable for analysis and decision support.

Feature Category	Example Variables	Data Type	Conceptual Role in Risk Modelling
Demographic	**Age, Gender**	**Numerical / Categorical**	**Provides baseline population characteristics and controls for demographic variability in mental health risk**
Academic	**Programme, Year of Study, Academic Performance**	**Categorical / Ordinal**	**Captures educational progression and attainment signals associated with academic stress and engagement**
Socioeconomic	**Income Level, Living Stability**	**Ordinal / Categorical**	**Reflects financial security and environmental stability, which influence stress exposure and coping capacity**
Psychological	**Stress, Depression, Anxiety**	**Numerical**	**Core proximal indicators of mental health vulnerability and primary drivers of high-risk classification**
Contextual (Aggregate)	**Conflict Exposure, Disaster Exposure**	**Ordinal / Binary**	**Encodes external stressors specific to fragile and conflict-affected settings, capturing background adversity**
Engineered	**Composite Stress Index, Contextual Stress Score**	**Numerical**	**Provides interpretable summarisation of multiple risk factors to support stable and explainable model behaviour**

Table 2. Feature categories, example variables, data types, and conceptual roles in mental health risk modelling.

This feature design enables the model to jointly account for individual vulnerability and contextual adversity, which is particularly important in FCAS higher-education settings.

4.4 Data Preprocessing and Class Imbalance Handling

values were replaced with the median value, while missing categorical values were completed using the most frequent category. Numerical features were then scaled using either min-max scaling or z-score standardisation, depending on the nature of the variable. Categorical variables were prepared according to their type. Nominal variables were converted using one-hot encoding, whereas ordered categories were encoded ordinally. Since the three risk groups were not equally represented, class imbalance was addressed within the training data using class weighting and oversampling. The original test distribution was kept unchanged to provide a more realistic evaluation.

4.5 Model Training, Evaluation, and Explainability

Two classifiers were used for benchmarking: Logistic Regression and Random Forest. The models were evaluated using stratified five-fold cross-validation so that each fold maintained a balanced representation of the three risk classes. To make the predictions more understandable, SHAP (Shapley Additive exPlanations) was applied after model training. This allowed the analysis to examine both overall feature importance and individual prediction explanations. Such interpretation is important in mental health-related applications, where model outputs should be transparent, explainable, and suitable for responsible use.

5. Results

5.1 Comparative Model Performance

Comparative results for the evaluated models are reported in Table 3. Logistic regression achieves the strongest overall performance, with an accuracy of 0.9350 and a macro-averaged F1-score of 0.8617, substantially outperforming the random forest baseline (accuracy = 0.8838, macro-F1 = 0.6541).

Model	Accuracy	Macro-F1	AUC (High-Risk, OvR)	Remarks
Logistic Regression	0.9350	0.8617	0.9969	Best overall performance, strong discrimination, well-calibrated probabilities, and high clinical utility
Random Forest	0.8838	0.6541	–	Lower balanced performance, reduced sensitivity to minority (high-risk) class

Table 3. Comparative performance of evaluated models under five-fold cross-validation.

Given its superior balanced performance (macro-F1), excellent discrimination for high-risk detection (AUC), and inherent interpretability, logistic regression was selected as the final model for detailed calibration, decision-analytic evaluation, and SHAP-based explanation (Figures 2–8).

5.2 High-Risk Detection Performance

The ability to identify students at high mental health risk is examined using a one-versus-rest ROC analysis. As shown in Figure 2, logistic regression achieves an AUC of 0.9969, indicating near-perfect discriminative capability.

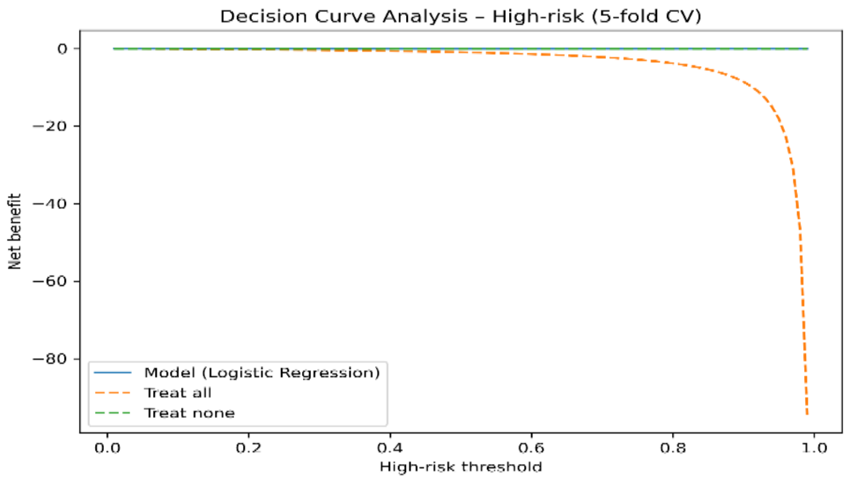

Figure 2. ROC curve for high-risk mental health detection using logistic regression (five-fold cross-validation).

The steep ascent of the ROC curve towards the upper-left corner indicates that high sensitivity can be achieved with minimal false-positive rates an important property in FCAS contexts where intervention resources are limited.

5.3 Calibration Performance

Figure 3 presents the calibration plot for high-risk predictions. The model demonstrates strong agreement between predicted probabilities and observed outcome frequencies, achieving a Brier score of 0.0241, which indicates low probabilistic error.

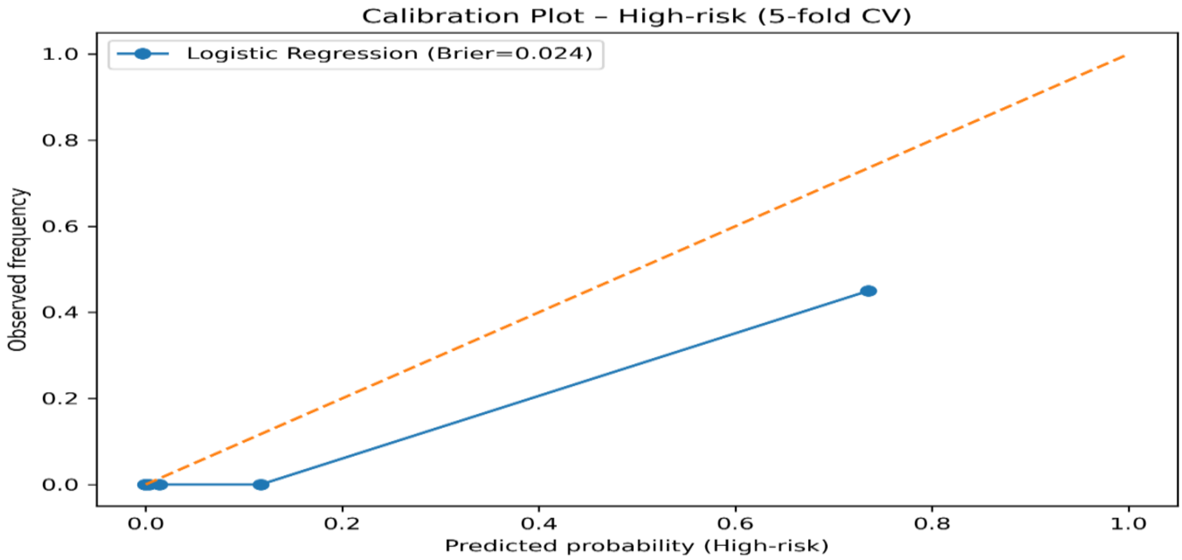

Figure 3. Calibration plot for high-risk predictions produced by logistic regression.

While slight under-confidence is observed at higher predicted probabilities, this behaviour is acceptable for early-warning systems, where conservative risk estimates reduce the likelihood of over-intervention.

5.4 Confusion Matrix Analysis

Class-specific performance is further illustrated in Figure 4, which presents the confusion matrix aggregated across cross-validation folds. The model shows strong separation between low- and high-risk categories, with no high-risk instances misclassified as low risk

.

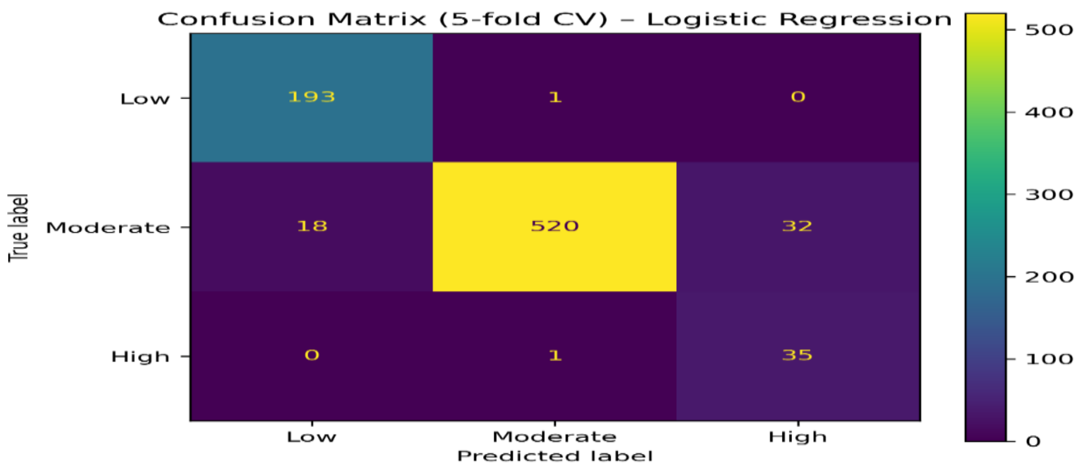

Figure 4. Confusion matrix for three-level mental health risk classification using logistic regression.

Most misclassifications occur between adjacent categories (low–moderate and moderate–high), reflecting the continuous nature of mental health risk rather than abrupt category boundaries.

5.5 Decision Curve Analysis

To move beyond discrimination and assess decision usefulness in practice, we conducted Decision Curve Analysis (DCA) for the high-risk class. DCA evaluates whether using the model to trigger follow-up support provides a net benefit across clinically plausible threshold probabilities, explicitly balancing the gains from correctly identifying high-risk students against the costs of unnecessary referrals. As shown in Figure 5, the logistic regression model achieves a consistently higher net benefit than both reference strategies treat-all and treat-none across a broad range of thresholds. In practical terms, this indicates that, for many reasonable operating points, the model would support a more efficient allocation of limited student-support resources than either intervening for everyone or not intervening at all (Figure 5).

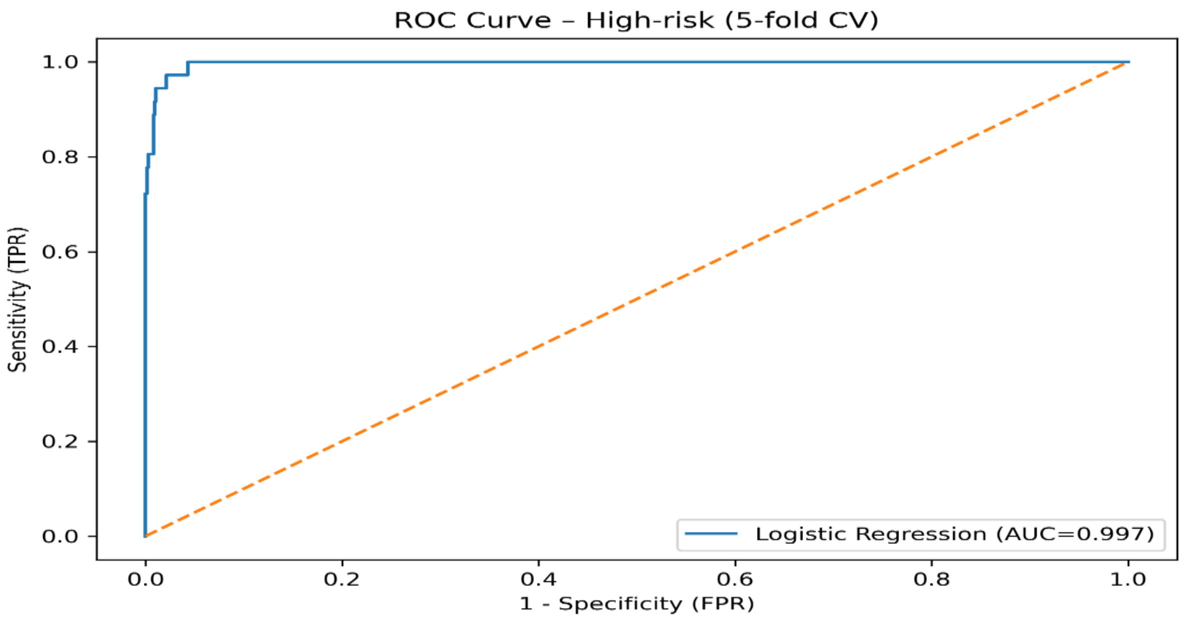

Figure 5. Decision curve analysis for high-risk mental health detection using logistic regression.

The DCA findings are coherent with the model's overall high-risk detection profile reported earlier. In particular, the panel summary in Figure 6 consolidates discrimination, calibration, and decision-analytic behaviour for high-risk prediction under five-fold cross-validation. The ROC panels in Figure 6 indicate strong discriminative performance for high-risk identification, while the calibration panels suggest generally reliable probability estimates (with low Brier error), and the DCA panels reiterate that the model's net benefit remains favourable relative to the treat-all and treat-none baselines over a wide threshold range (Figure 6). Taken together, Figures 5–6 indicate that the proposed model is not only accurate, but also *operationally meaningful* as a screening aid, particularly in resource-constrained university environments where decision thresholds may vary depending on staffing capacity and the availability of psychosocial services.

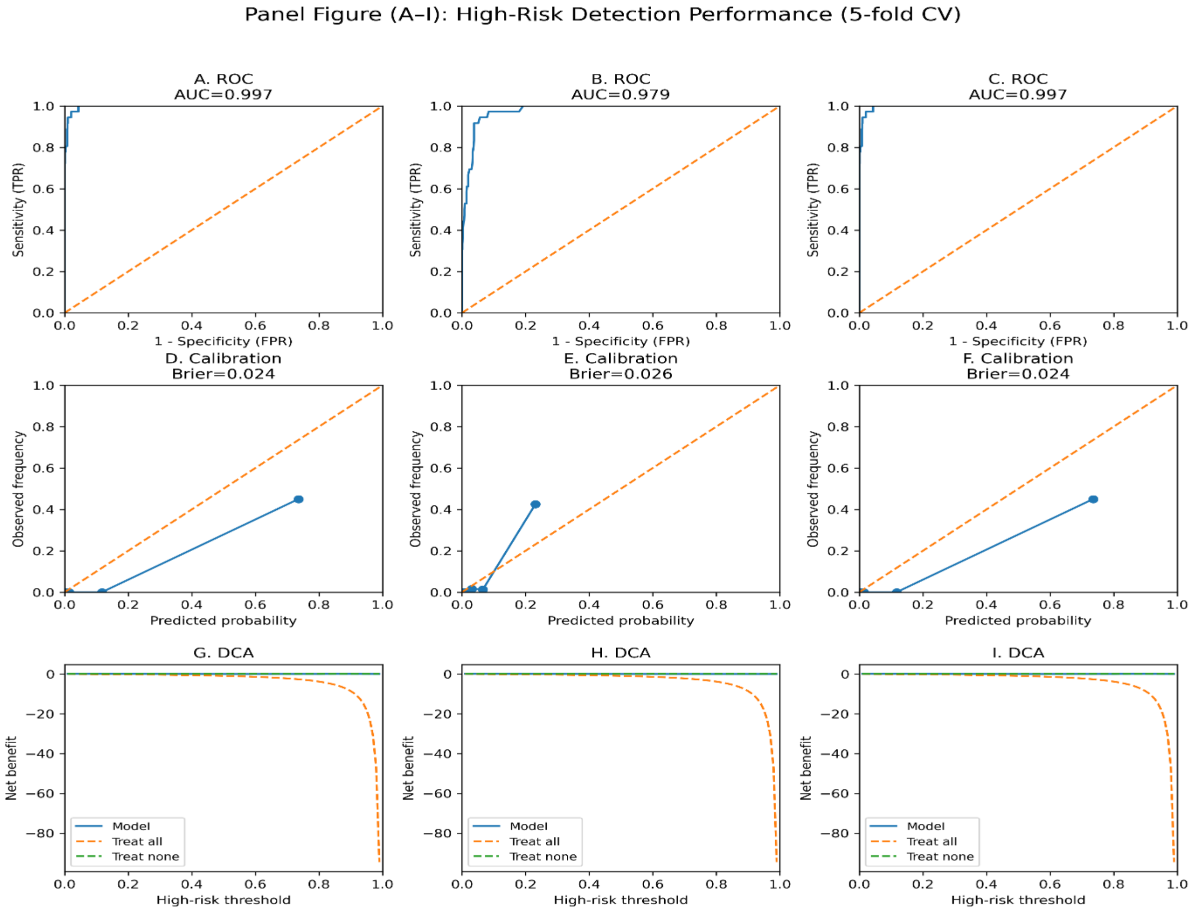

Figure 6. Panel summary of high-risk mental health detection performance under five-fold cross-validation.

5.6 Explainability Analysis Using SHAP

To enhance transparency and support deployment in sensitive educational contexts, the proposed framework incorporates SHAP (SHapley Additive exPlanations) to provide post-hoc interpretability of the high-risk mental health predictions. SHAP enables both global and local explanations by quantifying the marginal contribution of each feature to the model's output, thereby supporting trust, accountability, and actionable insight for decision-makers.

Figure **7** presents the global SHAP feature importance for high-risk prediction, measured as the mean absolute SHAP value across all instances. Psychological indicators dominate the model's decision process, with stress**,** depression**,** and anxiety emerging as the most influential predictors. Among these, *stress* exhibits the highest average contribution, followed closely by depression and anxiety, indicating that affective and emotional burden plays a central role in distinguishing high-risk students.

Beyond psychological factors, demographic and contextual variables such as age**,** conflict exposure**,** and income level contribute more modest but non-negligible effects. Academic performance shows a comparatively smaller influence, while gender and family history exhibit the lowest average impact. This hierarchy suggests that, within the studied setting, immediate psychological distress outweighs static demographic characteristics in driving high-risk mental health predictions.

Feature	Feature Category	Direction of Effect	Relative Importance (SHAP)	Interpretation for High-Risk Prediction
Stress	Psychological	Positive	Very High	Higher stress levels substantially increase the probability of being classified as high risk, making stress the strongest driver of model predictions.
Depression	Psychological	Positive	High	Elevated depression scores consistently contribute to increased predicted risk, reflecting its central role in mental health vulnerability.
Anxiety	Psychological	Positive	High	Anxiety shows a strong positive contribution, with higher values associated with increased risk, though with slightly lower impact than stress and depression.
Age	Demographic	Mixed	Low–Moderate	Age exhibits a modest and heterogeneous influence, suggesting non-linear or context-dependent effects on risk classification.
Conflict Exposure	Contextual	Positive	Low–Moderate	Exposure to conflict-related stressors contributes positively to risk, highlighting the influence of fragile and conflict-affected environments.
Income Level	Socioeconomic	Negative / Mixed	Low	Lower income levels are weakly associated with higher risk, though the effect is secondary to psychological indicators.
Academic Performance	Academic	Negative	Low	Lower academic performance slightly increases predicted risk, but its influence remains limited compared to psychological features.
Disaster Exposure	Contextual	Positive	Low	Disaster-related exposure shows a small positive contribution, indicating background environmental stress effects.
Family History	Demographic / Health	Positive	Very Low	Family mental health history has minimal influence in the presence of stronger psychological indicators.
Gender	Demographic	Neutral	Very Low	Gender contributes negligibly to predictions, suggesting the model does not rely on this sensitive attribute.

Table 4: SHAP-based explanation of key features for high-risk mental health prediction

Importantly, the presence of contextual stressors (e.g. conflict and disaster exposure) among the top-ranked contributors reinforces the relevance of modelling environmental adversity when analysing student mental health in fragile and conflict-affected settings.

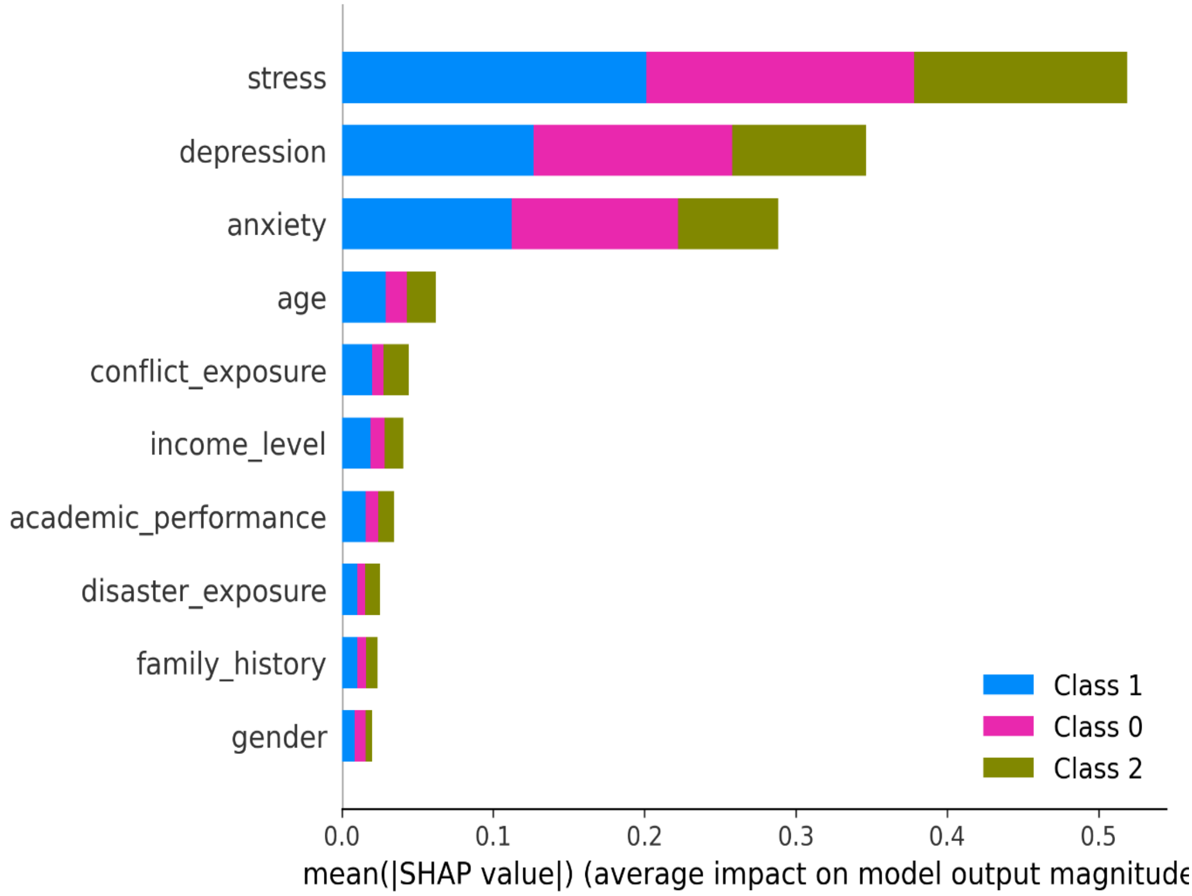

Figure 7**.** Global SHAP feature importance for high-risk mental health prediction, reported as mean absolute SHAP values across classes.

While global importance summarises average effects, Figure 8 provides a more granular view of feature-level contribution distributions, illustrating how individual observations influence model predictions through SHAP interaction values. The beeswarm-style visualisation shows that psychological variables not only dominate in magnitude but also display greater variability**,** indicating heterogeneous effects across students.

For example, *stress* and *depression* exhibit wide SHAP value spreads, suggesting that increases in these variables can substantially raise predicted risk for some students while exerting weaker effects for others. In contrast, demographic variables such as *age* and *gender* display narrower distributions centred close to zero, reflecting more stable and limited influence on the prediction outcome.

The interaction patterns further indicate that psychological and academic features interact weakly with demographic attributes, supporting the interpretation that risk elevation is primarily driven by internal psychological state**,** rather than demographic proxies alone. This behaviour is particularly important from an ethical perspective, as it reduces the likelihood of the model relying disproportionately on sensitive attributes when identifying high-risk individuals.

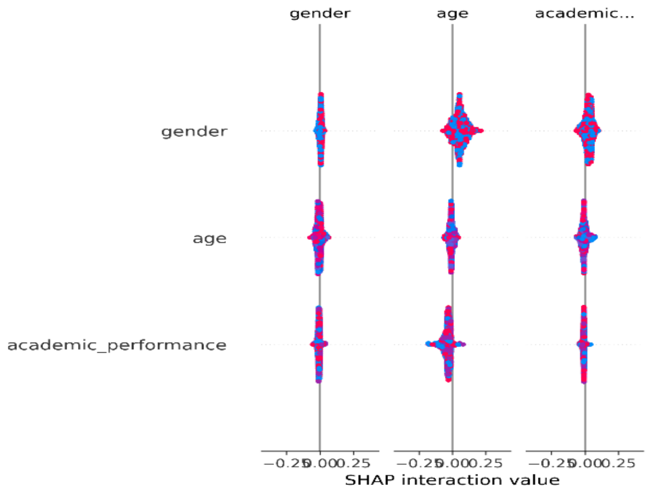

Figure 8**.** SHAP distribution and interaction effects for selected features.

Collectively, Figures 7–8 demonstrate that the proposed model's predictions are clinically and contextually interpretable**,** with dominant contributions arising from meaningful psychological constructs rather than opaque feature combinations. This transparency is essential for real-world adoption in university mental health screening, where model outputs must be explainable to counsellors, administrators, and ethical review boards.

By clearly identifying stress, depression, and anxiet*y* as primary drivers of high-risk classification while contextualising the secondary role of demographic and socioeconomic factors the SHAP analysis supports the use of the model as a decision-support tool rather than a black-box classifier. Such explainability is particularly critical in fragile and resource-constrained educational environments, where targeted and justifiable intervention strategies are required.

6. Discussion

The experimental findings demonstrate that interpretable and computationally lightweight models can deliver highly effective early mental health risk stratification when supported by coherent feature design and rigorous evaluation. In particular, logistic regression achieved the strongest overall performance under stratified five-fold cross-validation, reaching an accuracy of 0.9350 and a macro-averaged F1 score of 0.8617 (Table 3). This balanced performance is especially important in multi-class risk settings, where majority-class accuracy can mask poor detection of minority, high-risk cases. By comparison, the random forest baseline produced lower balanced performance (macro-F1 = 0.6541), reinforcing the value of simpler, well-regularised modelling in this context (Table 3).

A key practical objective of the proposed framework is reliable identification of students at high risk, where missed detection may carry the greatest safeguarding implications. The model showed near-perfect discriminative capability for high-risk identification, achieving a one-versus-rest AUC of 0.9969 (Figure 2). Importantly, the analysis extended beyond discrimination to examine probability reliability and decision usefulness, both of which are crucial for operational deployment. Calibration results indicate that predicted probabilities align closely with observed outcomes, with a low Brier score of 0.0241 (Figure 3), supporting the use of risk thresholds for triage and referral. Consistent with this, decision curve analysis demonstrated that the model yields a higher net benefit than treat-all and treat-none strategies across a wide range of threshold probabilities (Figure 5), indicating that the framework can support more efficient allocation of limited student-support resources. The integrated panel summary further consolidates this evidence by jointly presenting discrimination, calibration, and decision-analytic behaviour under cross-validation (Figure 6).

The confusion matrix provides additional insight into ethical and operational robustness. Misclassification patterns were largely concentrated between adjacent risk categories (low–moderate and moderate–high), which is conceptually consistent with the continuous nature of psychological distress. Crucially, high-risk cases were not misclassified as low risk, reducing the likelihood of overlooking students who may require timely support (Figure 4). This property is particularly salient in FCAS contexts, where institutional capacity for mental health services is constrained and conservative screening is often preferable to missed escalation.

Explainability results strengthen the case for responsible use in educational settings. SHAP analyses indicate that model predictions are primarily driven by psychologically meaningful indicators, with stress, depression, and anxiety emerging as dominant contributors to high-risk classification (Figures 7–8). Contextual variables related to FCAS adversity contribute secondary effects, while sensitive demographic attributes (e.g., gender) exhibit comparatively small influence, which is favourable from an ethical perspective. Together, these patterns suggest that the model's behaviour is largely aligned with clinically plausible constructs rather than demographic proxies, supporting transparency and stakeholder trust when the system is used as a screening aid.

Nevertheless, several limitations should be acknowledged. The current analysis is primarily cross-sectional, limiting the ability to capture risk trajectories over time and to evaluate how early-warning performance evolves across academic terms. In addition, some indicators are derived from self-reported measures, which may introduce response bias. Future work should therefore prioritise longitudinal modelling, external validation across independent institutions and cohorts, and evaluation of integration with real-world student-support workflows.

Ethical responsibility remains central to interpretation and potential translation. The study does not use identifiable personal data, medical records, or clinical diagnoses, instead, it relies on survey-informed indicators and aggregate contextual features, thereby reducing re-identification risk and supporting standard research governance principles (Ibrahim et al., 2025). The framework is intended as an early-warning decision-support approach rather than a diagnostic tool: model outputs should be treated as probabilistic risk indicators to inform proportionate triage and referral, and must not be used as a standalone basis for clinical judgement or disciplinary decision-making. Any real-world deployment would require formal ethical approval, institutional oversight, and informed consent, alongside clear governance on data access, retention, and escalation protocols, consistent with best practice in digital mental health research (Drira et al., 2024).

8. Limitations and Future Work

Several limitations should be acknowledged. First, although the proposed framework demonstrates strong predictive performance under stratified five-fold cross-validation (including high discrimination and favourable calibration), the evaluation remains confined to a single study dataset. As a result, the generalisability of the reported performance to other universities particularly across different FCAS contexts, cultures, and support infrastructures cannot be assumed. Future work should therefore prioritise external validation on independent cohorts and, where feasible, multi-institution benchmarking to assess robustness under context shift.

Second, while the feature set explicitly incorporates contextual adversity relevant to fragile and conflict-affected environments, some contextual indicators are necessarily represented at an aggregate or self-reported level and may not fully capture heterogeneity in individual exposure. Incorporating richer, institution-specific contextual variables potentially including temporally resolved indicators of disruption, displacement intensity, or service accessibility may improve sensitivity to local conditions and strengthen deployment realism.

Third, the current analysis is primarily cross-sectional, which limits the ability to model within-student trajectories and short-term changes in risk. Longitudinal designs, repeated measurements, and time-aware modelling would allow the framework to support dynamic early-warning use cases (monitoring risk progression across a semester) and improve alignment with intervention timing.

Finally, although SHAP-based explanations provide transparent and clinically plausible insights highlighting psychological indicators as dominant drivers of high-risk prediction explainability remains a technical account of model behaviour rather than a substitute for human judgement. Future research should include qualitative validation with counsellors, psychologists, and university support teams to assess whether the explanations are actionable, ethically appropriate, and aligned with institutional triage workflows. Such participatory evaluation would strengthen interpretive relevance and support responsible translation into practice.

9. Conclusion

This study developed and evaluated an explainable machine-learning framework for early mental health risk stratification among university students in fragile and conflict-affected settings (FCAS). By integrating survey-informed psychological indicators with academic, socioeconomic, and contextual stress representations, the framework moves beyond descriptive prevalence reporting towards proactive, interpretable, and operationally relevant risk prediction.

Empirically, the selected logistic regression model demonstrated strong and balanced predictive performance under stratified five-fold cross-validation, achieving an accuracy of 0.9350 and a macro-averaged F1 score of 0.8617. High-risk identification was particularly robust, with a one-versus-rest AUC of 0.9969, indicating near-perfect discriminative capability. Importantly, the model also produced reliable probabilistic estimates (Brier score = 0.0241) and exhibited favourable decision-analytic behaviour, yielding higher net benefit than treat-all and treat-none strategies across a wide range of plausible thresholds. Together, these results support the framework’s suitability as a screening-oriented decision aid in resource-constrained university environments.

Beyond performance, explainability analyses using SHAP provided transparent insight into model behaviour, showing that high-risk predictions are primarily driven by psychologically meaningful indicators most notably stress, depression, and anxiety with contextual variables contributing secondary effects and sensitive demographic attributes exerting minimal influence. This interpretability is central to responsible adoption in educational mental health settings, where accountability, trust, and safeguards against stigma are essential.

Overall, the study contributes an evidence-based foundation for context-aware, explainable early-warning support in FCAS higher education. Future work should prioritise (i) longitudinal modelling to capture risk trajectories, (ii) external validation across institutions and cohorts, and (iii) governance-aligned deployment studies that evaluate integration with counselling pathways, human oversight, and ethical safeguards in real-world practice.

References:

Alam Shuvo, T. (2025) 'Machine learning-based prediction of suicidal ideation, plans, and attempts among school-going adolescents', *Informatics and Health*, 2, pp. 143–157. https://doi.org/10.1016/j.infoh.2025.07.001

Arora, P. and Dahiya, S. (2025) 'Enhancing mental health diagnosis with DA-TabSVM: A multi-class hybrid approach for detecting depression and anxiety', *Procedia Computer Science*, 258, pp. 1348–1364. https://doi.org/10.1016/j.procs.2025.04.368

Biswas, J., Hasan, M.N., Gazi, M.S.R. and Rahman, M.M. (2025) 'Enhancing mental well-being: An artificial intelligence model for predicting mental disorders', *Array*, 26, 100417. https://doi.org/10.1016/j.array.2025.100417

Cai, B. and Wang, D. (2025) 'Prediction of psychological intervention for college students in digital entertainment media environment based on artificial intelligence and parallel computing algorithms', *Entertainment Computing*, 52, 100858. https://doi.org/10.1016/j.entcom.2024.100858

Blg et al. (2026) 'Machine Learning for Early Student Performance Prediction in E-Learning: A Systematic Literature Review', Preprint. https://journal.uob.edu.bh/items/0fd88300-b22c-485f-b190-d88740f3ac43

Fatahillah, A.F., Yauwrentius, N.A., Faiza, P.R., Hasani, M.F. and Maulina, A. (2025) 'Optimizing NLP for mental health analysis: Can DistilBERT match the performance of larger transformer models?', *Procedia Computer Science*, 269, pp. 1279–1288. https://doi.org/10.1016/j.procs.2025.09.069

Ibrahim, S.T., Li, M., Patel, J. and Katapally, T.R. (2025) 'Utilizing natural language processing for precision prevention of mental health disorders among youth: A systematic review', *Computers in Biology and Medicine*, 188, 109859. https://doi.org/10.1016/j.compbiomed.2025.109859

Kasereka, S.K., Tshibangu, K.N., Nyembo, M.M., Tshitenga, L.K., Muzindusi, M.K. and Kyamakyaf, K. (2025) 'Leveraging artificial intelligence for advancements in mental disorders: A short review', *Procedia Computer Science*, 257, pp. 676–683. https://doi.org/10.1016/j.procs.2025.03.087

Martínez-Romo, J., Araujo, L. and Reneses, B. (2025) 'Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports', *Computers in Biology and Medicine*, 186, 109701. https://doi.org/10.1016/j.compbiomed.2025.109701

Nath, M.D., Ahamed, M.K.U., Ahmed, O., Ahmed, T., Roy, S. and Uddin, M.N. (2025) 'Smart web interface for student mental health prediction using machine learning with blockchain technology', *Neuroscience Informatics*, 5, 100236. https://doi.org/10.1016/j.neuri.2025.100236

Suh, Y. and Yoo, J. (2025) 'Risk level prediction for problematic internet use: A digital health perspective', *Internet Interventions*, 41, 100863. https://doi.org/10.1016/j.invent.2025.100863

Al-Dekah, A.M. (2025) 'Mapping research trends and hotspots of artificial intelligence and machine learning in medicine and health within the Eastern Mediterranean region: A comprehensive bibliometric analysis', *Pharmacological Research – Reports*, 4, 100048. https://doi.org/10.1016/j.prerep.2025.100048

Hasan, M.E., Hasan, M.A., et al. (2025) 'Prevalence, associated factors, and machine learning-based prediction of depression, anxiety, and stress among university students: A cross-sectional study from Bangladesh', *Journal of Health, Population and Nutrition*, 44(1), 361. https://doi.org/10.1186/s41043-025-01095-8

Zhai, Y., Du, X. and Hou, Q. (2025) 'Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students', *Journal of Counseling & Development*, 103(1), pp. 45–58. https://doi.org/10.1002/jcad.12543

Yu, C., Kong, X., Yu, W., Ni, X., Chen, J. and Liao, X. (2025) 'Machine learning models for predicting the risk of depressive symptoms in Chinese college students', *Frontiers in Psychiatry*, 16, 1648585. https://doi.org/10.3389/fpsyt.2025.1648585

Opar, E., et al. (2024) 'Machine learning models to classify and predict depression in college students', *International Journal of Interactive Mobile Technologies*, 18(14), pp. 148–163. https://doi.org/10.3991/ijim.v18i14.48669

Drira, M., et al. (2024) 'Machine learning methods in student mental health: Ethical considerations, transparency and interpretability', *Applied Sciences*, 14(24), 11738. https://doi.org/10.3390/app142411738

Sahu, B., et al. (2023) 'Mental health prediction in students using data mining techniques', *Open Bioinformatics Journal*, 16, e187503622307140. https://doi.org/10.2174/18750362-v16-230720-2022-19

Daza, A. (2023) 'Systematic review of machine learning techniques to predict anxiety and stress', *Journal of Healthcare Informatics Research*, 9, pp. 365–387. https://doi.org/10.1007/s41666-023-00136-7

Campbell, A.T., et al. (2024) 'Explainable AI approaches for early detection of mental stress in university contexts', *Scientific Reports*, 15, 22171. https://doi.org/10.1038/s41598-025-22171-3

El Morr, C., et al. (2024) 'Predictive machine learning models for assessing mental health outcomes in diverse student populations', *Journal of Primary Care & Community Health*, 15, 21501319241235588. https://doi.org/10.1177/21501319241235588

Baba, A. and Bunji, K. (2023) 'Prediction of mental health problems using annual student health surveys: A machine learning approach', *BMC Psychiatry*, 23, 567. https://doi.org/10.1186/s12888-023-04567-4

g training sets. Ai & Society, 36(4), 1105-1116.