MindCare: An Interpretable Multimodal Framework for Early Student Mental-Wellness Risk Assessment Using NLP and Bayesian Inference

Anonymous authors
Paper under double-blind review

Abstract

Student mental-wellness assessment involves heterogeneous evidence, including self-report, journal text, and optional behavioural indicators. We present MindCare, a lightweight evidence-fusion framework that combines a 12-question instrument incorporating DASS-21-derived and CHU9D-derived items, rule-assisted linguistic features, and normalized sleep/activity indicators within an interpretable Bayesian Network. Journal text is processed using phrase matching, domain-specific keywords, VADER sentiment, and TextBlob polarity/subjectivity. The network produces posterior estimates for anxiety, depression, and stress. Evaluation comprises functional testing on 20 controlled cases, external evaluation of four NLP-derived features with conventional classifiers, and Bayesian sensitivity and evidence-source ablation analyses. The internal functional test achieved 90.0% accuracy, whereas the external seven-class dataset yielded macro F1 scores of 21.61%, 21.01%, and 26.56% for Gaussian Naive Bayes, Logistic Regression, and Random Forest, respectively. Sensitivity and ablation experiments showed directionally consistent evidence responses. These results establish technical feasibility and interpretability of the implemented evidence-fusion mechanism, but do not establish clinical validity or population-level generalization because the internal test is small, external labels do not match the Bayesian outputs, and CPTs are predefined rather than empirically calibrated.

Keywords: mental health assessment, natural language processing, Bayesian Network, evidence fusion, student wellness, sentiment analysis, explainable AI.

1 Introduction

Mental-health difficulties among students can be influenced by academic, financial, and social pressures, while barriers to help-seeking can delay access to support (Kessler et al., 2007). Digital assessment systems therefore increasingly combine structured questionnaires with computational analysis of unstructured text and other behavioural signals. Recent transformer-based approaches demonstrate strong performance on mental-health text classification, while multimodal systems combine text with audio, visual, or behavioural information (Khan et al., 2025; Goyal et al., 2025; Wan et al., 2025; Jiang, 2026). Such systems, however, can require substantially more computation or richer sensing than lightweight educational deployments.

MindCare instead treats journal language as an evidence source rather than as a standalone diagnostic classifier. It combines a structured questionnaire, journal-derived linguistic features, and optional normalized sleep/activity indicators in a Bayesian Network whose conditional relationships are explicit. The objective is therefore not to outperform large language or transformer models on text classification, but to provide a transparent mechanism for combining heterogeneous evidence and propagating uncertainty to anxiety, depression, and stress risk estimates.

Our contributions are: (1) a lightweight multimodal evidence-fusion architecture; (2) a rule-assisted linguistic pipeline using phrase matching, keywords, VADER, and TextBlob; (3) an explicit

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Bayesian Network for probabilistic risk inference; (4) sensitivity and evidence-source ablation analyses; and (5) external evaluation of the lightweight NLP feature representation using conventional machine-learning baselines.

2 Related Work

NLP-based mental-health research has highlighted both predictive potential and persistent challenges in annotation, data quality, external validation, and real-world deployment (Ibrahim et al., 2025; Montejo-Raez et al., 2024). Khan et al. reported 97% accuracy using RoBERTa-Large and ELECTRA on a seven-class student mental-health classification task (Khan et al., 2025). Wan et al. similarly used a RoBERTa-GRU framework for college-student depression detection, while Jiang explored interpretable cross-attention transformer models (Wan et al., 2025; Jiang, 2026). These studies demonstrate the predictive capacity of modern representation learning but differ from MindCare in computational and inferential design.

Multimodal systems provide richer evidence. MindLift combines behavioural tracking, audio, facial expression, and text using CNN, RNN, and transformer components (Goyal et al., 2025); MhAFN combines text and audio with multi-head attention (Agrawal & Iyer, 2025). Other work has investigated LLaMA-based psychometric analysis and hybrid educational agents (Agrawal & Iyer, 2026b;a). These approaches illustrate the value of heterogeneous evidence but generally use more complex learned representations than the feature-based pipeline considered here.

Digital self-guided interventions provide a complementary perspective, with Bakker et al. evaluating fully self-guided CBT applications for depression and anxiety (Bakker et al., 2018). Yasmin et al. investigated explainable machine learning for university-student mental-health prediction (Yasmin et al., 2026). MindCare differs by making evidence relationships explicit in the Bayesian Network itself rather than relying primarily on post-hoc explanations. The resulting research gap is a lightweight framework in which structured questionnaire evidence, interpretable linguistic signals, and optional behavioural indicators are combined through an explicit probabilistic inference structure.

3 MindCare Framework

3.1 Evidence Sources and Architecture

MindCare supports three evidence modes. Journal mode captures free-text entries through the web interface. Questionnaire mode collects responses through a 12-question instrument incorporating DASS-21-derived and CHU9D-derived items (Lovibond & Lovibond, 1995; Rowen et al., 2012), producing structured indicators related to anxiety, mood, fatigue, sleep, and functioning. Behavioural mode optionally supplies sleep efficiency, REM percentage, resting-heart-rate change, and step-count change relative to a seven-day baseline. Derived features are timestamped and stored for the assessment workflow; raw journal content is processed in memory and discarded according to the implemented pipeline.

All continuous evidence variables are normalized and then discretized into the three categorical states required by the Bayesian Network. Questionnaire-derived anxiety, depression, and stress scores use the implemented thresholds $\leq 7/8$–$14/ > 14$, $\leq 13/14$–$20/ > 20$, and $\leq 14/15$–$25/ > 25$, respectively. Normalized sleep and activity factors use the inverse mapping ≥ 66, 33–65, and < 33. Journal polarity uses ≤ -0.10, -0.10 to < 0.10, and ≥ 0.10. These are model-specific computational states and are not presented as official clinical severity categories.

3.2 Linguistic Feature Extraction

The journal pipeline uses three stages. First, phrase matching captures multi-word expressions such as "can't stop worrying" and "don't see the point." Second, domain-specific keyword matching detects depression-, anxiety-, and positive-emotion terms. Third, VADER compound sentiment provides a fallback sentiment signal (Hutto & Gilbert, 2014). For this fallback, scores ≥ 0.15 are classified as Positive, scores ≤ -0.15 as Depression, and intermediate scores as Neutral. TextBlob polarity and subjectivity are retained as additional continuous features. The four features used in

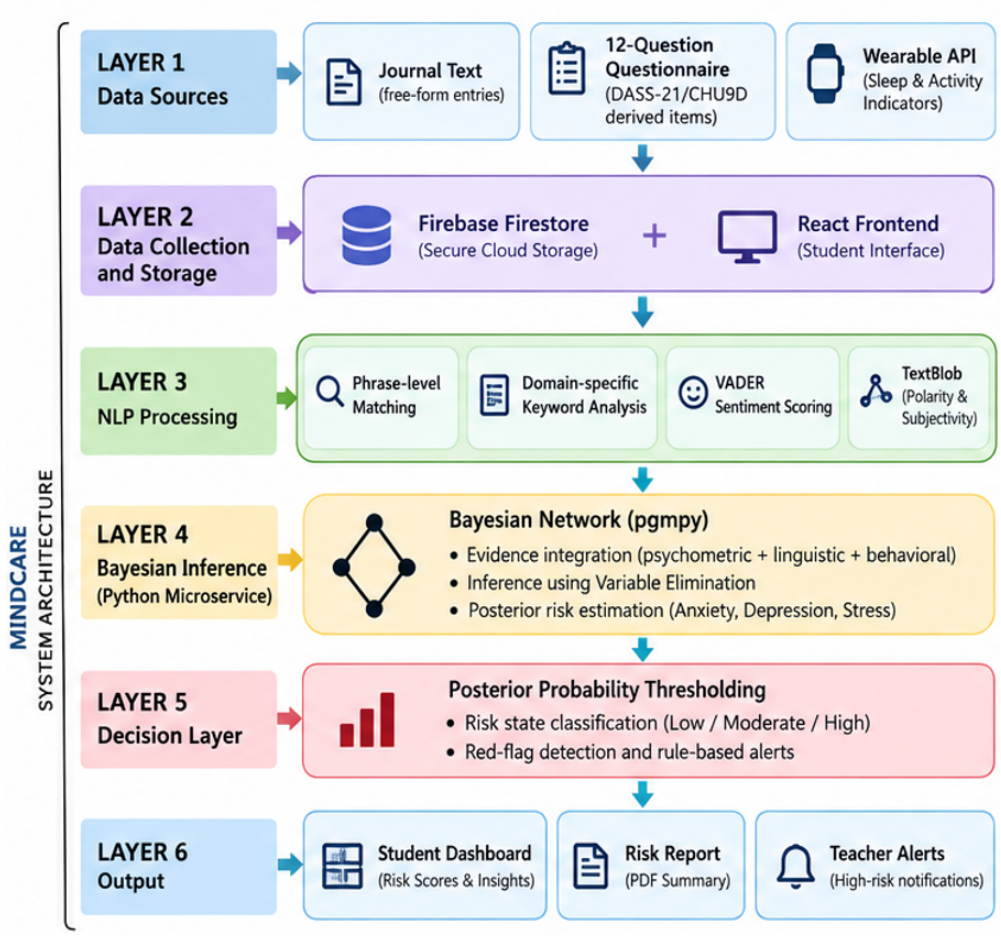


Figure 1: MindCare evidence-processing architecture: data sources, collection, NLP processing, Bayesian inference, decision, and output.

the external baseline experiment are TextBlob polarity, TextBlob subjectivity, VADER compound score, and detected keyword count.

3.3 Bayesian Network

The Bayesian Network is implemented with pgmpy and Variable Elimination. Its dependency structure is: Questionnaire Anxiety$\times Sleep \rightarrow AnxietyRisk$, $QuestionnaireDepression \times JournalSentiment \rightarrow DepressionRisk$, $QuestionnaireStress \times Activity \rightarrow StressRisk$. All nodes have Low, Moderate, and High states. The CPTs were predefined from literature-informed directional assumptions rather than learned from a sufficiently large empirical dataset. For example, poorer sleep is represented as evidence associated with greater anxiety risk, while negative linguistic cues are represented as evidence associated with greater depression risk (Coppersmith et al., 2015). The numerical CPT entries operationalize the model structure and should not be interpreted as empirical clinical probabilities.

For reproducibility, the implemented example assigns P(A=H$|\ Q_A = H, S = P) = 0.80$, $P(D = H \mid Q_D = H, J = N) = 0.80$, where A, D denote anxiety and depression risk, Q denotes questionnaire evidence, S sleep, and J journal sentiment. These values are model parameters, not estimates reported by the cited studies.

Posterior distributions are converted to an interpretable 0–100 score using

$$R = 20P(L) + 55P(M) + 90P(H). \tag{1}$$

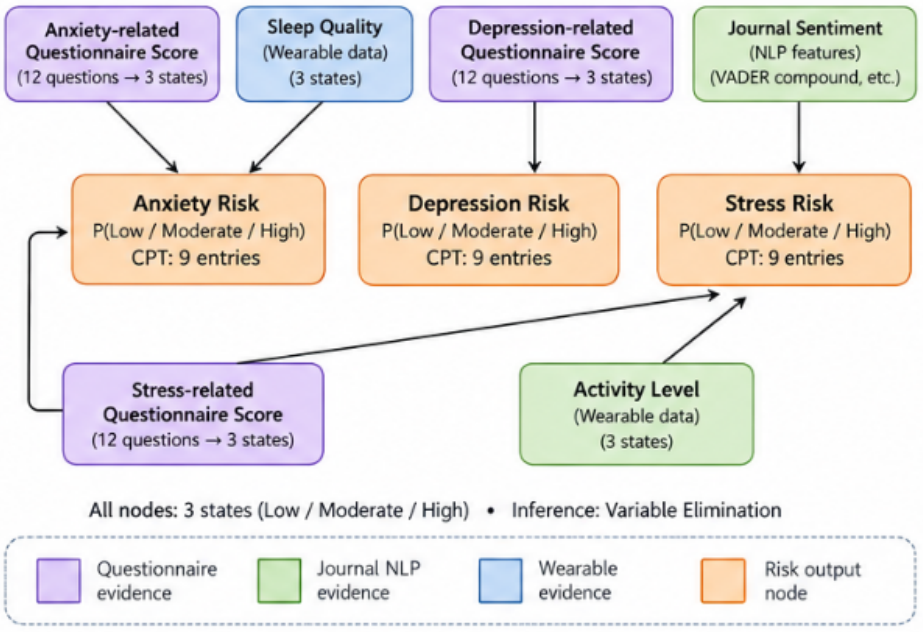


Figure 2: MindCare Bayesian Network. Six evidence nodes propagate through predefined conditional relationships to three risk outputs.

Thus, the system preserves the posterior distribution while providing a compact score for the dashboard.

4 Experimental Setup

We evaluate four aspects of the implementation. First, functional NLP validation uses 20 manually constructed journal-like cases labelled Depression, Anxiety, Positive, or Neutral. Because these cases were designed using vocabulary represented in the rule lexicon, this experiment tests implementation behaviour rather than generalization. Second, the extracted NLP representation is evaluated on an independent external dataset after removal of missing statements, conflicting labels, and duplicates. The cleaned data are divided using a stratified 80/20 train-test split. Gaussian Naive Bayes, Logistic Regression, and Random Forest use the same four NLP-derived features. The external dataset contains seven status categories that do not directly correspond to MindCare's three Bayesian outputs; consequently, this experiment evaluates the representational capability of the linguistic features rather than the complete risk model.

Third, sensitivity analysis varies one evidence variable at a time while holding other inputs fixed and converts the resulting posterior distributions using Eq. 1. Fourth, evidence-source ablation compares questionnaire-only, questionnaire-plus-journal, questionnaire-plus-behaviour, and full multimodal configurations on one representative high-risk profile. The ablation measures evidence contribution, not predictive accuracy.

5 Results

5.1 Functional NLP Validation

The internal functional evaluation achieved 90.0% accuracy, 89.2% weighted F1, 91.7% weighted precision, and 90.0% weighted recall. Per-class results are shown in Table 1. Anxiety achieved an F1-score of 100%, while Depression and Positive achieved 90.9%; Neutral achieved 75.0% because two Neutral instances were misclassified.

Table 1: Internal functional NLP validation (20 cases).

Class	Precision	Recall	F1
Depression	83.3	100.0	90.9
Anxiety	100.0	100.0	100.0
Positive	83.3	100.0	90.9
Neutral	100.0	60.0	75.0
Weighted	91.7	90.0	89.2

Table 2: External evaluation of the four lightweight NLP features.

Model	Acc.	M. Prec.	M. Rec.	M. F1
Gaussian NB	50.24	29.17	24.56	21.61
Logistic Regression	35.98	23.14	25.02	21.01
Random Forest	43.72	26.47	26.91	26.56

5.2 External Evaluation of NLP Features

The external experiment shows a substantially different picture from the controlled internal test. Gaussian Naive Bayes achieved the highest accuracy at 50.24%, while Random Forest achieved the highest macro F1 at 26.56%. The low macro-level scores indicate that the four lightweight linguistic features alone are insufficient for robust seven-class mental-health status classification, particularly across minority classes.

These results support the intended role of NLP in MindCare as a lightweight evidence extractor rather than a standalone diagnostic classifier. They also provide an important guard against interpreting the 20-case internal result as evidence of generalization.

5.3 Bayesian Sensitivity

Varying individual evidence variables produced directionally consistent changes in the associated risk estimates. Increasing questionnaire-derived anxiety, depression, and stress evidence increased the corresponding risk scores, while reducing sleep/activity factors or journal polarity increased the associated risks. The plateaus at some values arise from three-state discretization. These experiments therefore provide structural and directional validation of the implemented inference mechanism rather than clinical calibration.

5.4 Evidence-Source Ablation

On the representative high-risk profile, adding journal evidence increased depression risk from 72.5 to 81.2, while behavioural evidence increased anxiety and stress risk. This behaviour follows the predefined network structure. Because the experiment uses one profile and predefined CPTs, it demonstrates evidence contribution rather than population-level predictive improvement.

6 Discussion and Limitations

The experiments establish MindCare as a technically interpretable evidence-fusion framework rather than a clinically validated diagnostic system. The external baseline results show that lightweight linguistic features alone are insufficient for robust multi-class classification, supporting their use as complementary evidence. Sensitivity analysis confirms that the Bayesian inference mechanism responds directionally to questionnaire, linguistic, and behavioural evidence, while ablation shows that different evidence sources affect outputs according to the predefined dependency structure.

Several limitations constrain interpretation. The internal NLP evaluation contains only 20 manually constructed cases and therefore cannot establish population-level generalization. The external dataset contains seven status categories that do not directly correspond to the three Bayesian risk

Table 3: Bayesian sensitivity analysis.

Evidence	Input variation	Risk variation
Questionnaire anxiety	5→10→15→20	30.5→44.5→67.2→67.2
Sleep factor	80→60→30→10	30.5→37.5→48.0→48.0
Questionnaire depression	5→14→18→25	30.5→46.2→46.2→65.5
Journal polarity	+0.50→0→-0.20→-0.70	30.5→39.2→53.2→53.2
Questionnaire stress	5→15→20→30	32.2→41.7→41.7→65.5
Activity factor	80→60→30→10	32.2→39.2→51.5→51.5

Table 4: Evidence-source ablation on a representative high-risk profile.

Configuration	Anxiety	Depression	Stress
Questionnaire only	74.2	72.5	72.5
Questionnaire + Journal	74.2	81.2	72.5
Questionnaire + Behaviour	81.2	72.5	79.5
Full multimodal	81.2	81.2	79.5

outputs. The CPTs are predefined from literature-informed assumptions rather than learned from a sufficiently large longitudinal clinical dataset, so the resulting scores are not clinically calibrated. The behavioural component currently uses normalized sleep and activity indicators, and the evaluation does not include prospective validation, demographic subgroup analysis, calibration assessment, or comparison with clinically established screening outcomes. These limitations define the present work as a technical feasibility and interpretability study. Future work should focus on larger longitudinal datasets, expert-informed and data-driven CPT estimation, probability calibration, aligned external validation, subgroup robustness, and prospective evaluation.

7 Conclusion

MindCare provides a lightweight mechanism for combining questionnaire-derived, linguistic, and optional behavioural evidence through an interpretable Bayesian Network. The internal test verifies functional operation of the rule-assisted NLP component, while the external experiment demonstrates that its lightweight linguistic representation is not sufficient as a standalone seven-class classifier. Bayesian sensitivity and ablation experiments demonstrate interpretable changes in posterior risk when evidence sources are varied. The current results support technical feasibility of the evidence-fusion mechanism, but not clinical validity. Empirical parameter learning, calibration, aligned external validation, and longitudinal evaluation remain necessary before any clinical deployment claim can be made.

AI Use Statement

Generative AI tools were used during preparation for language editing, structural refinement, and limited programming/debugging assistance. The authors reviewed AI-assisted text and suggestions against the implemented system and reported results and take responsibility for the final content, claims, analyses, and artifacts.

Ethics Statement

MindCare is presented as a technical feasibility and interpretable evidence-fusion framework, not a clinically validated diagnostic system. It should not replace professional assessment. Future work involving real student data will require appropriate ethical oversight, informed consent, privacy safeguards, demographic fairness analysis, and domain-expert validation. The implemented pipeline processes journal content in memory and retains derived linguistic features rather than raw journal text.

Reproducibility Statement

The paper specifies the evidence sources, linguistic features, discretization thresholds, Bayesian dependency structure, CPT assumptions, posterior score mapping, external feature set, data-cleaning procedure, train-test split, sensitivity protocol, and ablation configurations. These details provide the basis for reproducing the reported analyses. Source code and supplementary implementation material should be supplied anonymously where permitted by the submission system.

References

S. Agrawal and R. Iyer. Mhafn: A multi-head attention-based fusion network for personality prediction from multimodal text and audio data. In *Proceedings of the 2025 2nd International Conference on Cognitive Robotics and Intelligent Systems*, pp. 216–221, 2025. doi: 10.1109/ICC-ROBINS64345.2025.11086199.

S. Agrawal and R. Iyer. Ai-based educational interactive agents using hybrid query classification. *International Journal of Software Engineering and Knowledge Engineering*, 2026a. doi: 10.1142/S0218194025500925.

S. Agrawal and R. Iyer. Psychometric analysis using llama based circular position embedding and gated residual auto-encoder. *International Journal of Information Technology*, 2026b. doi: 10.1007/s41870-025-03072-0.

M. Bakker et al. Effectiveness of fully self-guided digital cbt interventions for depression and anxiety: An rct. *JMIR Mental Health*, 5(2):e10770, 2018.

G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 31–39, 2015. doi: 10.3115/v1/W15-1204.

S. Goyal, R. Dutta, S. Dev, K. N. Raju, and M. W. Bhatt. Mindlift: Ai-powered mental health assessment for students. *Neuroscience Informatics*, 5(2):100208, 2025. doi: 10.1016/j.neuri.2025.100208.

C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, pp. 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

S. T. Ibrahim, M. Li, J. Patel, and T. R. Katapally. Utilizing natural language processing for precision prevention of mental health disorders among youth: A systematic review. *Computers in Biology and Medicine*, 188:109859, 2025. doi: 10.1016/j.compbiomed.2025.109859.

D. Jiang. Predicting student mental health through entropy-based features and interpretable cross-attention transformer networks. *PLOS ONE*, 21(4):e0347294, 2026. doi: 10.1371/journal.pone.0347294.

Ronald C. Kessler et al. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the world mental health survey initiative. *World Psychiatry*, 6(3):168–176, 2007.

H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam. Analyzing student mental health with roberta-large: a sentiment analysis and data analytics approach. *Frontiers in Big Data*, 8:1615788, 2025. doi: 10.3389/fdata.2025.1615788.

P. F. Lovibond and S. H. Lovibond. The structure of negative emotional states: Comparison of the dass with beck inventories. *Behaviour Research and Therapy*, 33:335–343, 1995.

A. Montejo-Raez, M. D. Molina-Gonzalez, S. M. Jimenez-Zafra, M. A. Garcia-Cumbreras, and L. J. Garcia-Lopez. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654, 2024. doi: 10.1016/j.cosrev.2024.100654.

D. Rowen et al. Estimating a preference-based index from the chu9d. *Health and Quality of Life Outcomes*, 10:57, 2012.

Q. Wan, Y. Pan, and S. Zakeri. Analyzing depression in college students using nlp and transformer models: Implications for career and educational counseling. *Brain and Behavior*, 15(9):e70828, 2025. doi: 10.1002/brb3.70828.

M. N. Yasmin, M. M. Rahman, N. Y. Ahmed, et al. An explainable machine learning framework for analyzing and predicting mental health problems among university students in bangladesh. *Scientific Reports*, 16:23179, 2026. doi: 10.1038/s41598-026-52351-8.