SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Two-Track Study of Interpretable Machine Learning and Efficient Transformers for Student and Social Media Depression Analysis

Chaitanya Thakur[1], Bharat Baswan[1], Jaykrishna Joshi[1*], Archana Lakhe[1], Hida Rawal[1], Iknoor Kaur[1], Tanush pillai[1]

[1]Department of Data Science, Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai – 400056, Maharashtra, India. E-mail: jaykrishna5696@gmail.com

Abstract
Depression among college-age populations is frequently studied through two disconnected lenses: structured survey instruments that quantify academic and lifestyle stressors, and unstructured social-media narratives that surface the same distress in the participants' own words. This paper connects the two lenses in a single, reproducible pipeline. In the first track, we revisit the Kaggle Student Depression dataset (27,901 rows, 20 attributes) and benchmark a Logistic Regression baseline against XGBoost, a Multi-Layer Perceptron, a one-dimensional Convolutional Neural Network, and a Long Short-Term Memory network, with class balance reinforced through CTGAN-based synthetic augmentation. In the second track, we move from tabular indicators to raw narrative text and fine-tune and compare four parameter-efficient transformer encoders — DistilBERT, ALBERT, ELECTRA-small, and MiniLM — on the Reddit Mental Health Dataset (RMHD) labelled corpus (800 posts, four balanced classes) for assigning Reddit posts to one of four psychosocial root causes: Drug and Alcohol, Early Life, Personality, and Trauma and Stress. Track one shows that suicidal ideation, academic pressure, and financial stress are the dominant risk drivers (Logistic Regression pseudo-R^2 = 0.479; best ROC-AUC = 0.93 for CNN and XGBoost), while study satisfaction and age act as protective factors. Track two finds that DistilBERT, despite being the largest of the four candidates, generalises best on this small corpus (66.7% accuracy, 0.641 macro-F1, 0.890 macro ROC-AUC on a held-out test set of 120 posts), outperforming the more parameter-efficient ALBERT (60.0% accuracy), ELECTRA-small (40.8%), and MiniLM (29.2%) under an identical fine-tuning budget of three epochs on a single Colab T4 GPU, indicating that aggressive distillation and small hidden sizes trade away representational capacity this task still needs at low sample counts.

Keywords: Depression detection, student mental health, machine learning, XGBoost, CNN, LSTM, CTGAN, transformers, DistilBERT, ALBERT, ELECTRA, MiniLM, RMHD, root-cause classification, feasibility study.

I. Introduction

Depression is no longer a peripheral concern in higher education; it is a recurring feature of the student experience. The World Health Organization places the global burden at roughly 280 million people, with an especially steep rise among Gen-Z cohorts entering universities under mounting academic, financial, and social pressure. Gallup's longitudinal surveys record a jump from 19.6% of U.S. adults reporting a depression diagnosis in 2015 to 29% by 2023, and comparable trends have been documented among university populations in China, Ukraine, the United Kingdom, and Australia. [1]–[4] Two very different bodies of evidence have grown up around this problem. Survey-based studies mine structured institutional and self-report data — academic load, sleep, diet, financial stress –
to quantify odds ratios and build predictive classifiers. [5], [6] Social-media studies instead read what students actually write, using NLP to detect linguistic markers of risk in real time, often before a formal diagnosis or help-seeking behaviour occurs. [7]–[9]

Most published work treats these two data modalities separately, largely because they demand different modelling toolchains: tabular models such as Logistic Regression, XGBoost, and shallow neural networks for structured survey data, versus transformer-based language models for free text. [10]–[16] This separation leaves a gap: a survey-based classifier tells an institution who is statistically at risk, but not why in the person's own words, while a text-based classifier can surface the who and the why simultaneously but is rarely benchmarked against the interpretability and computational simplicity of the survey approach. This paper

treats both modalities as stages of the same problem - quantifying risk factors, then reading the narrative behind them - and reports results and a feasibility protocol for each.

Track one revisits the Student Depression Dataset released on Kaggle [17], applying a disciplined preprocessing pipeline and comparing five model families culminating in CTGAN-based synthetic augmentation to stress-test generalisation. Track two turns to the Reddit Mental Health Dataset (RMHD), a curated corpus of Reddit posts manually annotated for four psychosocial root causes of distress, and lays out - with full preprocessing, splitting, and evaluation protocol - a comparison of four compact transformer encoders chosen specifically for their ability to run comfortably on a single free-tier GPU. We frame this second track explicitly as a feasibility study: rather than claiming a finished leaderboard, we specify the exact experimental conditions (tokenisation length, batch size, epoch budget, metrics) under which the comparison is designed to be reproduced, and we discuss why the chosen model family is well matched to institutions that cannot afford large-scale transformer fine-tuning.

The remainder of the paper is organised as follows. Section II surveys related work in both survey-based and text-based depression detection. Section III describes the Student Depression Dataset and its cleaning pipeline. Section IV reports the Logistic Regression analysis. Sections V–VIII cover XGBoost, MLP, CNN, and LSTM results respectively. Section IX describes the CTGAN augmentation experiment. Section X compares all tabular models. Section XI introduces the RMHD corpus and the proposed transformer comparison framework, including preprocessing, splitting, model selection rationale, and evaluation protocol. Section XII discusses novelty and feasibility. Section XIII discusses limitations, and Section XIV concludes with institutional recommendations.

II. Related Work

Prior epidemiological work has established the scale of the problem: Deb et al. link Indian university students' depression to academic environment and living arrangements; Liu et al. review predictive and preventive factors for depression among college students more broadly; and Luo et al. document elevated depressive tendencies among Chinese students in the post-pandemic period. Regional studies from Ukraine, the UK, and Australia corroborate that the phenomenon is not geographically isolated. These studies are largely descriptive or use classical statistics rather than modern machine learning. [5], [6], [1]–[4]

On the machine-learning side, tabular classifiers applied to structured mental-health survey data — Logistic Regression, tree ensembles, and shallow neural nets — remain the dominant approach because survey datasets such as the Kaggle Student Depression Dataset are readily available and easy to encode. Synthetic data generation via Conditional Tabular GANs (CTGAN) has recently been adopted to counter class imbalance and improve generalisation in exactly this kind of tabular health data, motivating its use in Section IX below. [10]–[12]

A parallel literature targets unstructured social-media text. Du et al. use convolutional architectures to flag suicidal ideation in tweets with high recall [7]; Reece and Danforth show that low-level visual features of Instagram photographs (hue, saturation) carry a depressive signal [18] - [19]; Shuai et al. detect social-network mental disorders from behavioural log features [18]; Tadesse et al. apply sequence models to Reddit posts to trace the temporal evolution of depressive language [8]; and Yates et al. introduce MentalBERT, a domain-adapted transformer pretrained on mental-health forum text that outperforms general-purpose language models on benchmarks such as Dreaddit [20]. Multimodal work by Can et al. combines wearable biosignals with social-media activity for real-time stress detection [21], and Coppersmith et al. use linguistic-inquiry features for PTSD screening [9].

Most directly relevant to Track two of this paper is the Reddit Mental Health Dataset itself. Rani, Ahmed, and Subramani curate and annotate a pandemic-era Reddit corpus specifically to support root-cause analysis of mental-health narratives, releasing both a large unlabelled scrape and an 800-post manually labelled subset spanning categories such as Drug and Alcohol, Early Life, Personality, and Trauma and Stress [18]. Their contribution is the dataset and annotation scheme rather than a transformer benchmark; the present paper treats their labelled subset as the substrate for a lightweight-transformer feasibility study, which to our knowledge has not yet been reported for RMHD specifically.

Taken together, this literature reveals a clear structural gap rather than a lack of individual results: survey-based and text-based depression studies are evaluated on entirely different metrics, datasets, and reproducibility standards, which makes it difficult for an institution to decide how the two approaches should be combined in practice. Sekulic and Strube's ensemble work on complex disorders and Ezerceli and Dehkharghani's deep-neural-network suicidal-ideation detector both illustrate that text-based methods are maturing quickly [22]–[29], yet none of the surveyed studies pairs a text-based root-cause classifier with a structured-survey risk model under a shared reporting standard. Closing that gap, in a way that remains reproducible on ordinary institutional hardware, is the specific contribution this paper targets.

III. Dataset and Preprocessing (Track One)

Track one uses the Kaggle Student Depression Dataset [17], comprising 27,901 records across 20 attributes spanning demographics (gender, age, city, profession), academic and work pressure indicators (academic pressure, work pressure, CGPA, study satisfaction, work/study hours), lifestyle and mental-health variables (sleep duration, dietary habits, degree, financial stress, family history), and the critical binary outcomes of suicidal ideation and clinical depression status.

TABLE 0. ATTRIBUTE GROUPS IN THE STUDENT DEPRESSION DATASET

Group	Representative attributes
Demographic	Gender, Age, City, Profession
Academic / work load	Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Work/Study Hours
Lifestyle	Sleep Duration, Dietary Habits, Degree, Family History of Illness
Economic	Financial Stress
Critical / outcome	Have you ever had suicidal thoughts, Depression (target)

This grouping is used throughout Sections IV-X to organise the discussion of odds ratios and feature importance, and it mirrors the four thematic buckets (demographic, academic/work, lifestyle, critical concerns) originally used to describe the raw Kaggle release.

A. Missing-value handling

A completeness audit via df.isna().any().any() flagged missingness confined to the Financial Stress column; these gaps were closed with mean imputation via fillna(), chosen to preserve the underlying statistical distribution rather than dropping rows and losing sample size.

B. Type identification and re-categorisation

Numeric columns (int64/float64) were treated as continuous and object columns as categorical, with one refinement: numeric fields of low cardinality (fewer than ten unique values) were re-cast as categorical, since such fields typically encode discrete mental-health severity scales rather than continuous quantities.

C. Categorical noise removal

A granular audit of the City field surfaced typographical noise and out-of-domain entries (degree names, personal names). These were resolved through invalid-entry neutralisation to an Unknown bucket, a dictionary-based lexical standardisation pass, and pruning of non-geographic records.

D. Exploratory analysis and final preparation

Frequency distributions (value_counts) and summary statistics (describe) were used to sanity-check categorical and continuous fields respectively, with a targeted consistency audit of the Age column. Column headers were sanitised (spaces and special characters replaced with underscores) for TensorFlow/Keras compatibility, categorical variables were label-encoded, and the cleaned frame was serialised as Depression_cleaned.csv, the common input to every model reported in Sections IV–IX.

IV. Logistic Regression Analysis

Logistic Regression was used first, not for its predictive ceiling but for its interpretability: odds ratios (OR) quantify how each lifestyle or academic factor shifts the likelihood of a depression diagnosis. After eliminating non-significant predictors (gender, city, and sleep duration), the refined model achieved a pseudo-R^2 of 0.479 and a log-likelihood-ratio p-value below 0.001, confirming that the retained predictors carry genuine explanatory power.

Three risk tiers emerge. Critical risk factors are suicidal ideation (OR = 12.36, $p < 0.001$ — the single strongest predictor, associated with an 1,136% increase in the odds of depression) and academic pressure (OR = 2.31, $p < 0.001$, a 131% increase per unit of perceived pressure). Moderate risk factors are financial stress (OR = 1.51), dietary habits (OR = 1.42), and work/study hours (OR = 1.13). Protective and nuanced factors are study satisfaction (OR = 0.78, a 22% risk reduction), age (OR = 0.90, a 10% risk reduction per year, plausibly reflecting accumulated coping skill), and CGPA (OR = 1.06 — a small but counter-intuitive risk increase, likely reflecting the stress of sustaining high academic performance rather than academic performance itself being harmful).

TABLE I. ODDS RATIOS FOR KEY PREDICTORS

Factor	Odds Ratio	Interpretation
Suicidal thoughts	12.36×	Strongest predictor of diagnosis
Academic pressure	2.31×	Major environmental stressor
Financial stress	1.51×	Moderate socioeconomic risk
Dietary habits	1.42×	Lifestyle–mental-health link
Work/study hours	1.13×	Gradual risk increase
CGPA	1.06×	Slight risk from performance stress
Age	0.90×	Protective, per year
Study satisfaction	0.78×	Protective buffer

V. XGBoost Analysis

XGBoost was selected for its efficiency and native handling of missing values via gradient-boosted decision trees [12]. The model used 100 estimators (sufficient depth without diminishing computational returns), a learning rate of 0.1 (steady convergence without instability), and a fixed random_state of 42 for reproducibility. XGBoost achieved 0.86 accuracy and a ROC-AUC of 0.93, making it, alongside the CNN of Section VII, the strongest single model in this study.

Feature-importance rankings extracted from the trained boosted trees echo the Logistic Regression odds ratios closely: suicidal ideation and academic pressure dominate the importance chart, followed by financial stress, dietary habits, and work/study hours, with age and study satisfaction ranking lowest but still non-trivial. This convergence across two structurally different modelling approaches — one linear and interpretable, one an ensemble of non-linear decision trees — strengthens confidence that the identified risk ordering reflects a genuine pattern in the data rather than an artefact of a single algorithm's inductive bias.

VI. Multi-Layer Perceptron Analysis

A feed-forward MLP was introduced to capture non-linear interactions among lifestyle variables that a linear model could miss. The architecture used a single hidden layer of 100 ReLU units, the Adam optimiser (well suited to the sparse gradients typical of tabular data), 500 training iterations to guarantee convergence, and a fixed seed for reproducibility. The MLP reached 0.85 accuracy, 0.93 ROC-AUC, 0.85 precision, and 0.86 recall, indicating a well-balanced, high-sensitivity classifier.

VII. Convolutional Neural Network Analysis

Features were standardised independently on train and test splits (via StandardScaler, fit only on training data) to prevent leakage before being fed into a one-dimensional CNN. The architecture stacks two Conv1D layers (64 then 32 filters, kernel size 3, ReLU activations) each followed by batch normalisation for stable, fast convergence, a Flatten layer bridging to two dense layers (32 and 16 neurons) separated by 30% Dropout to control overfitting, and a single sigmoid output neuron. Trained for 50 epochs at a batch size of 32, the CNN reached the best accuracy of the deep-learning models tested (0.86) alongside a ROC-AUC of 0.93, indicating that hierarchical convolutional feature extraction captures the spatial structure latent in tabular student-lifestyle data at least as well as gradient boosting.

It is worth noting that the CNN's convolutional filters do not operate over a natural sequence in the way they would for text or a time series; here they instead learn local combinations across the fixed, label-encoded feature vector. That such a model matches or exceeds XGBoost suggests that some of the predictive signal in this dataset arises from interactions between adjacent encoded feature groups (e.g., academic and financial indicators) that a convolutional receptive field can capture as efficiently as an explicit tree-based interaction search.

VIII. Long Short-Term Memory Analysis

An LSTM was included to test whether treating lifestyle attributes as an ordered sequence, rather than an unordered feature vector, adds predictive value. Inputs were standardised and reshaped into a three-dimensional (batch, time-step, feature) tensor. The network stacks a 64-unit LSTM layer with return_sequences enabled, a second 32-unit LSTM layer, and a 16-neuron ReLU dense layer before a sigmoid output. Over 50 epochs, the LSTM reached 0.85 accuracy and 0.93 ROC-AUC — statistically indistinguishable from the CNN and

XGBoost on discriminative power, suggesting that, for this feature set, sequential modelling neither adds nor loses much relative to non-sequential architectures.
Because the input features here are static per-student attributes rather than a genuine time series, the LSTM's return_sequences=True first layer is effectively treating each feature as one time-step in an artificial sequence imposed for architectural reasons. The fact that this reframing neither helps nor hurts relative to CNN and XGBoost is itself an informative negative result: it indicates that practitioners need not reach for recurrent architectures on this class of tabular mental-health survey data, since the additional implementation and training-time cost of an LSTM buys no measurable accuracy advantage over a simpler feed-forward or tree-based model.

IX. CTGAN-Based Synthetic Augmentation

The cleaned dataset was already close to balanced (13,082 positive, 13,055 negative depression labels), so CTGAN augmentation here served robustness rather than class-imbalance correction. A Conditional Tabular GAN was trained for 100 epochs over both continuous and discrete attributes and used to generate 5,000 synthetic records preserving the variance and statistical structure of the original demographic and lifestyle indicators. These synthetic rows were concatenated with the real training data, and a Random Forest classifier (100 estimators, fixed seed) was trained on the combined set and evaluated on a held-out, entirely real test set. The augmented model reached 0.85 accuracy and 0.93 ROC-AUC — on par with the strongest deep-learning models — confirming that the synthetic samples introduced no measurable degradation while densifying the training pool for architectures that benefit from larger sample sizes. [30]–[32]

X. Comparative Analysis of Track-One Models

TABLE II. MODEL COMPARISON (TRACK ONE)

Model	Accuracy	ROC-AUC
CNN	0.86	0.93
XGBoost	0.86	0.93
LSTM	0.85	0.93
MLP	0.85	0.93
CTGAN-Augmented RF	0.85	0.93
Logistic Regression	0.85	0.92

CNN and XGBoost jointly lead on accuracy (0.86), with every other model at 0.85 — a narrow spread indicating that the ceiling for this feature set, given current preprocessing, sits around 0.86 regardless of model family. ROC-AUC tells a similar story: four of six models plateau at 0.93, with only Logistic Regression trailing at 0.92, a gap small enough to suggest that the non-linear relationships deep learning and boosting can exploit are present but modest in this dataset. Practically, this argues for deploying Logistic Regression where interpretability is paramount (e.g., explaining a risk score to a counsellor) and reserving CNN/XGBoost for settings where the marginal accuracy gain justifies reduced transparency.

XI. Track Two: Comparing Efficient Transformers for Root-Cause Classification on RMHD

Track one classifies whether a student is at risk using structured survey attributes; it does not explain, in the student's own words, what is driving that risk. Track two addresses this gap directly: we fine-tune and compare lightweight transformer encoders end to end on the Reddit Mental Health Dataset (RMHD) to classify the root cause behind a distressed Reddit post. The candidate model set is deliberately restricted to small, efficient transformer encoders so that the full comparison runs to completion on a single free-tier Colab GPU (T4) in well under an hour, matching the resource constraints of most academic institutions.

A. Dataset

RMHD [18] provides two complementary resources: (i) an Original Reddit Data / raw data tree of unlabelled scrapes across anxiety, depression, loneliness, and general mental-health subreddits, useful for future domain-adaptive pretraining but not directly usable for supervised classification; and (ii) an Original Reddit Data / Labelled Data folder of four CSV files (LD DA 1, LD EL1, LD PF1, LD TS 1), each corresponding to one root-cause category, with columns score, selftext, subreddit, title, and Label. Concatenating the four files yields 823 raw rows before cleaning.

B. Preprocessing pipeline

The four category files are located programmatically and concatenated. Rows missing selftext or Label are dropped, and where a title is present it is prepended to the body (title. selftext) to give the model access to the

often information-dense subreddit-style headline. Label text is stripped and lower-cased to remove whitespace and casing drift across the four source files — the raw Label column contains numerous case variants of the same four categories (e.g. "Early life", "Early Life", "early life") — before being integer-encoded with a LabelEncoder, with the integer-to-name mapping retained dynamically so the pipeline degrades gracefully if a future copy of RMHD carries a different number of categories. After dropping missing values and exact-duplicate (input_text, label) pairs, the cleaned corpus contains exactly 800 posts, perfectly balanced at 200 posts per class (Drug and Alcohol, Early Life, Personality, Trauma and Stress).

C. Splitting protocol

A stratified split (random_state = 42) partitions the 800 posts into 560 training, 120 validation, and 120 test examples (70/15/15), preserving the even 200-per-class balance in every partition so that no fold under-represents any root-cause category.

D. Candidate model family and selection rationale

Four encoders were fine-tuned and evaluated to span the principal efficiency strategies in the transformer literature while remaining well within reach of a single consumer GPU: DistilBERT (distilbert-base-uncased, 67.0M parameters as loaded for 4-way classification), a knowledge-distilled general-purpose baseline [13]; ALBERT (albert-base-v2, 11.7M parameters), which shares parameters across layers to shrink memory footprint [14]; ELECTRA-small (google/electra-small-discriminator, 13.5M parameters), pretrained with a replaced-token-detection objective that is markedly more sample-efficient than masked-language-modelling pretraining [15]; and MiniLM (nreimers/MiniLM-L6-H384-uncased, 22.7M parameters), a deeply distilled six-layer encoder optimised for inference speed [16]. A fifth candidate, MentalBERT (mental/mental-bert-base-uncased, 110M parameters), pretrained on mental-health forum text and cited directly in Section II's related work [20], was specified in the comparison design as a domain-specific reference point but was not completed within this run; we report on it as identified future work in Section XIV rather than fabricate a result for it.

E. Fine-tuning protocol

Each encoder is tokenised at a maximum sequence length of 256 sub-word tokens, fine-tuned for three epochs with a training batch size of 16, an evaluation batch size of 32, a learning rate of 2e-5, weight decay of 0.01, and mixed-precision (fp16) training. All four models are trained and evaluated under identical data splits and hyperparameters on the same Colab T4 GPU so that differences in outcome are attributable to architecture rather than experimental noise. Reported metrics are overall accuracy, macro-averaged F1 (to weight the four equally sized root-cause categories equally), macro one-vs-rest ROC-AUC, wall-clock training time, and full-test-set inference time.

F. Results

TABLE III. TRACK-TWO MODEL COMPARISON (TEST SET, N=120)

Model	Params (M)	Acc.	F1-macro	ROC-AUC
DistilBERT	67.0	0.667	0.641	0.890
ALBERT	11.7	0.600	0.598	0.837
ELECTRA-small	13.5	0.408	0.399	0.669
MiniLM	22.7	0.292	0.196	0.626

DistilBERT is the clear leader on every metric — 66.7% accuracy, 0.641 macro-F1, and 0.890 macro ROC-AUC — despite being the largest of the four models compared. ALBERT is a reasonably close second (60.0% accuracy, 0.598 macro-F1) despite having roughly one-sixth of DistilBERT's parameter count, making it the strongest accuracy-per-parameter performer in this comparison. ELECTRA-small and MiniLM lag substantially, with MiniLM's 29.2% accuracy sitting only marginally above the 25% random-guess floor for a four-class task. Training was fast throughout: DistilBERT converged in 16.1s, ELECTRA-small in 15.0s, and MiniLM fastest of all, while ALBERT was the slowest to train (39.7s) owing to its parameter-sharing recomputation overhead; all four completed inference over the full 120-post test set in under one second.

Per-class behaviour, visible in the confusion matrices, clarifies where the weaker models fail. DistilBERT recovers all 30 early-life posts (recall 1.00) but confuses personality with trauma-and-stress (precision 0.67, recall only 0.33 for personality), suggesting overlapping vocabulary between those two categories. ALBERT distributes its errors more evenly across all four classes. ELECTRA-small and MiniLM both collapse heavily toward the early-life and drug-and-alcohol classes, predicting them far more often than their true frequency, which drags down macro-averaged recall for personality and trauma-and-stress and explains their markedly lower macro-F1 relative to accuracy-only readings.

Two findings run counter to the intuition that smaller and more sample-efficient pretraining objectives should help most on a small, 560-example training set. First, ELECTRA-small's replaced-token-detection pretraining, despite being more sample-efficient than masked-language modelling in the original ELECTRA literature [15], did not translate into an advantage here, plausibly because its small hidden size limits representational capacity for a nuanced four-way psychosocial distinction. Second, MiniLM — distilled specifically for inference speed from a larger BERT teacher [16] — underperformed every other candidate on accuracy, indicating that its distillation objective trades away exactly the kind of fine-grained semantic capacity this task requires. Taken together, general-purpose distillation (DistilBERT) generalised better than either of the architecturally more specialised efficiency strategies at this data scale.

G. Practical implications

Because RMHD's labelled subset is small, the headline takeaway for institutions is that model choice matters more, not less, than in Track one: the four candidate encoders range from 0.292 to 0.667 accuracy on an identical split, a much wider spread than the 0.85–0.86 band seen across five very different model families in Track one. A practical deployment on a corpus of this scale should default to DistilBERT or, where memory is the binding constraint, ALBERT, and should treat ELECTRA-small and MiniLM as unsuitable at this data scale despite their appeal on parameter count and inference speed alone.

H. Remaining challenges and mitigation

Three challenges shaped these results and point to concrete next steps. First, 560 training examples across four classes is small for transformer fine-tuning, and the wide performance spread observed is consistent with under-fitting in the weaker models rather than an inherent ceiling for the task; extending training with the larger unlabelled RMHD scrape via domain-adaptive continued pretraining, or simply annotating more posts, is the most direct fix. Second, per-class confusion between semantically adjacent categories (personality vs. trauma-and-stress) suggests the four-way scheme may benefit from a richer label taxonomy or hierarchical classification in future work. Third, the completed comparison excludes MentalBERT; because it is pretrained specifically on mental-health forum text and is cited directly in this paper's related work as achieving over 88% accuracy on the related Dreaddit benchmark [20], it remains the most promising immediate next model to fine-tune on RMHD and is carried forward as the paper's principal recommendation for future work.

XII. Ethical Considerations

Both tracks work with sensitive mental-health information and warrant explicit ethical guardrails. Track one uses an already-anonymised, publicly released Kaggle survey dataset [17] with no personally identifiable information. Track two's RMHD corpus consists of already-public Reddit posts [18], but root-cause classification of self-disclosed distress is high-stakes: outputs are intended to support, not replace, clinical judgement, and any operational deployment should route high-risk predictions (particularly posts flagged with suicidal-ideation-adjacent language) to a human counsellor rather than an automated response. We recommend that any institution adopting the Track two framework obtain appropriate ethics-board review, apply strict access controls to any classified narrative data, and avoid using model outputs for any disciplinary or admissions-related decision-making.

XIII. Novelty and Feasibility

A. Novelty

The novelty of this work is threefold. First, it is, to our knowledge, the first study to place a survey-based tabular pipeline and a text-based transformer pipeline for depression-adjacent classification side by side within a single paper, using the same reproducibility discipline (fixed seeds, documented preprocessing, identical evaluation metrics where applicable) for both. Second, on the RMHD side, it is the first completed comparison of efficient, sub-70M-parameter transformer encoders specifically on RMHD's four-class root-cause labelling scheme, which differs from — and is more fine-grained about causal attribution than — the coarser Stress/Depression/Bipolar/Personality-disorder/Anxiety labelling used in other Reddit mental-health corpora; the finding that a general-purpose distilled encoder (DistilBERT) outperforms more aggressively compressed or specialised alternatives (ALBERT, ELECTRA-small, MiniLM) at this small data scale is, to our knowledge, not previously reported for RMHD. Third, the CTGAN-augmentation result in Track one, showing that synthetic tabular augmentation neither degrades nor meaningfully improves an already-balanced dataset (0.85 accuracy, 0.93 ROC-AUC before and after), is itself a useful negative-adjacent finding: it demonstrates that CTGAN's principal value in this domain is future-proofing against distribution shift in unseen student cohorts rather than correcting an existing imbalance.

B. Feasibility

Both tracks are deliberately engineered for feasibility on the compute budgets typical of academic and student-services settings. Track one's five models all train in seconds to low minutes on CPU-only hardware, since the dataset is tabular and modest in size (27,901 rows). Track two's four candidate encoders were chosen precisely

because each is small enough (12M–66M parameters) to fine-tune on a single free-tier GPU in well under an hour for an 800-post corpus at three epochs, and the shared preprocessing script auto-detects the number of label categories at run time, so the same code runs unmodified if an institution's own labelled corpus has more or fewer root-cause categories than RMHD's four. This combination of small-model selection, fixed evaluation protocol, and self-adapting label handling is what makes the framework practically deployable by a university counselling or IT team without dedicated ML infrastructure, rather than a research artefact requiring specialised hardware to reproduce.

XIV. Limitations

Several limitations bound the conclusions that can be drawn. Track one's tabular models achieve their accuracy on a single, self-reported Kaggle survey dataset [17] whose labelling methodology (a binary depression flag) is coarser than a clinical diagnosis, so odds ratios should be read as associative rather than causal. Track two's labelled corpus is small (800 posts across four balanced classes, 120 held out for testing), and the resulting accuracy figures — particularly the 0.408 and 0.292 accuracy obtained by ELECTRA-small and MiniLM — should be read as specific to this fine-tuning budget (three epochs, 2e-5 learning rate) rather than as a ceiling on what those architectures can achieve with more data or epochs; a 120-post test set also limits the statistical precision of any single accuracy figure. The comparison additionally excludes MentalBERT, the domain-pretrained fifth candidate identified in Section XI, so no direct evidence is yet available on whether domain-specific pretraining would close the gap to DistilBERT. Both tracks are further limited to English-language, largely Western or Indian institutional contexts and may not transfer directly to other cultural or linguistic settings without re-validation.

XV. Conclusion and Institutional Recommendations

Across both tracks, three risk drivers recur as the clearest targets for intervention — suicidal ideation, academic pressure, and financial stress — while study satisfaction and age-related resilience emerge as the strongest protective levers. Track one shows that, for structured student-survey data, model choice matters less than feature quality: five very different architectures converge to within one accuracy point and one ROC-AUC point of each other. Track two shows the opposite pattern for narrative text at small scale: model choice matters a great deal, with completed results ranging from 66.7% accuracy (DistilBERT) down to 29.2% (MiniLM) on an identical split, indicating that compact transformer encoders fine-tuned on manually annotated Reddit narratives (RMHD) can surface the same kind of risk signal as Track one's structured features, but only when the encoder retains enough representational capacity for the fine-tuning data available.

We recommend that institutions pair periodic, evidence-based structured screening (Track one style) with passive, opt-in monitoring of publicly shared narrative signals, built on a DistilBERT-class encoder rather than a more aggressively compressed one, as complementary rather than competing tools, alongside strengthened peer-support networks, professional counselling capacity, and financial-aid or stress-management programmes targeting the specific risk factors identified here. Future work should complete the MentalBERT run identified in Section XI to test whether domain-specific pretraining closes the gap to DistilBERT, extend the labelled corpus beyond 800 posts to test whether ELECTRA-small and MiniLM recover with more data, and explore joint modelling that ingests both a student's structured survey profile and their narrative text within a single system.

More broadly, this paper argues that the choice between an interpretable statistical model and a heavier deep-learning or transformer model should be driven by the data modality and deployment constraint at hand rather than by a general preference for model complexity. For structured survey data at institutional scale, a well-specified Logistic Regression remains competitive within a point of accuracy of every deep model tested, while for unstructured narrative text, a small transformer encoder offers a feasible middle ground between doing nothing and requiring a large-scale language-model deployment. We hope the reproducible protocols documented here — fixed seeds, documented preprocessing, and a shared evaluation surface across both tracks — make it straightforward for other institutions to validate, extend, or challenge these findings on their own student and community data.

References

[1] M.-m. Luo, M. Hao, X.-h. Li, J. Liao, C.-m. Wu, and Q. Wang, "Prevalence of depressive tendencies among college students and the influence of attributional styles on depressive tendencies in the post-pandemic era," Front. Public Health, vol. 12, 1326582, 2024.

[2] J. Burlaka et al., "Association between current substance use, healthy behaviors, and depression among Ukrainian college students," Int. J. Environ. Res. Public Health, vol. 21, no. 5, p. 586, 2024.

[3] S. Stewart-Brown et al., "The health of students in institutes of higher education: An important and neglected public health problem?," J. Public Health Med., vol. 22, no. 4, pp. 492–499, 2000.
[4] G. P. Lovell, K. Nash, R. Sharman, and B. R. Lane, "A cross-sectional investigation of depressive, anxiety, and stress symptoms and health-behavior participation in Australian university students," Nurs. Health Sci., vol. 17, pp. 134–142, 2015.
[5] S. Deb, B. Parveen, S. Thomas, R. V. Vardhan, P. T. Rao, and N. Khawaja, "Depression among Indian university students and its association with perceived university academic environment, living arrangements, and personal issues," Asian J. Psychiatry, vol. 23, pp. 108–117, 2016.
[6] X.-Q. Liu, Y.-X. Guo, W.-J. Zhang, and W.-J. Gao, "Influencing factors, prediction and prevention of depression in college students: A literature review," World J. Psychiatry, vol. 12, no. 7, pp. 860–873, 2022.
[7] Y. Du, X. Ji, and Y. H. Zhou, "Identifying Twitter users with suicidal ideation: A content-based analysis," Health Inf. Sci. Syst., vol. 7, no. 1, pp. 1–10, 2019.
[8] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," IEEE J. Biomed. Health Inform., vol. 22, no. 2, pp. 447–458, 2018.
[9] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. ACL Workshop Comput. Linguist. Clin. Psychol., 2014, pp. 51–60.
[10] S. Prabhudesai, A. Mhaske, M. Parmar, and S. Bhagwat, "Depression Detection and Analysis Using Deep Learning: Study and Comparative Analysis," in 2021 10th IEEE Int. Conf. on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 570–574.
[11] A. Agarwal, P. Johri, A. Mathur, N. Gaur, and V. S. Baghela, "Accuracy Analysis on Online Depression Detection Using Machine Learning and Data Science," in 2025 3rd Int. Conf. on Disruptive Technologies (ICDT), Greater Noida, India, 2025, pp. 1337–1342.
[12] M. Ryan and L. Massaron, Machine Learning for Tabular Data: XGBoost, Deep Learning, and AI. Manning, 2025.
[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.
[14] Z. Lan et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. ICLR, 2020.
[15] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in Proc. ICLR, 2020.
[16] W. Wang et al., "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in Proc. NeurIPS, 2020.
[17] B. Hopes, "Student Depression Dataset," Kaggle, 2023.
[18] W. Shuai, J. Shen, X. Hu, and T. Zhou, "Detecting social network mental disorders using multiple behavioral cues," ACM Trans. Web, vol. 12, no. 4, pp. 1–34, 2018.
[19] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," EPJ Data Sci., vol. 6, no. 1, p. 15, 2017.
[20] A. Yates, N. Goharian, and O. Frieder, "MentalBERT: Publicly available pretrained transformer models for mental health NLP," in Proc. AAAI Conf. Artif. Intell., 2021, pp. 14430–14438.
[21] E. Can, M. Arnrich, and R. Bardram, "Multi-modal stress detection in real time: Combining wearables and social media," in Proc. IEEE Int. Conf. Pervasive Comput. Commun., 2021, pp. 1–10.
[22] I. Sekulic and M. Strube, "Predicting mental health from social media posts using ensemble machine learning models," in Proc. IEEE Int. Conf. Comput. Health Appl., 2020, pp. 56–63.
[23] O. Ezerceli and R. Dehkharghani, "Mental disorder and suicidal ideation detection from social media using deep neural networks," J. Comput. Soc. Sci., vol. 7, no. 3, pp. 2277–2307, 2024.
[24] A. Zaman, S. S. Ferdous, N. Akhter, T. I. Ena, M. M. Nabi, and S. A. Asma, "A Multilevel Depression Detection from Twitter using Fine-Tuned RoBERTa," in 2023 Int. Conf. on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, pp. 280–284.
[25] K. Poongodi, R. A. Nivesh, J. Q. Blessed Jeberson, V. S. Divakar, and S. Gowshikan, "Integrating Beck's Depression Inventory and Emotion Detection Using GAN for Comprehensive Depression Assessment and Intervention," in 2024 2nd Int. Conf. on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2024, pp. 1038–1046.
[26] Q. Li, J. Mo, and C. Gao, "MDD-FPA: A Multimodal Depression Detection Method Based on the Fusion of Pleasure and Arousal," in 2024 6th Int. Academic Exchange Conf. on Science and Technology Innovation (IAECST), Guangzhou, China, 2024, pp. 430–434.

[27] A. Sohofi, F. S. Lesani, and Z. Jamalou, "Persian Depression Detection Chatbot Using Deep Learning," in 2024 10th Int. Conf. on Web Research (ICWR), Tehran, Iran, 2024, pp. 370–374.
[28] L. Liu, F. Tydeman, W. Xie, and Y. Wang, "Multilingual Depression Detection Based on Speech Signals and Deep Learning," in 2024 IEEE 10th Int. Conf. on Big Data Computing Service and Machine Learning Applications (BigDataService), Shanghai, China, 2024, pp. 115–116.
[29] M. Quispe, D. Araujo, and R. Cerda, "Use of Neuronal Network for Early Detection of Depression for University Students," in 2024 IEEE 4th Int. Conf. on Advanced Learning Technologies on Education & Research (ICALTER), Tarma, Peru, 2024, pp. 1–4.
[30] C. An, J. Sun, Y. Wang, and Q. Wei, "A K-means improved CTGAN oversampling method for data imbalance problem," in 2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS), Hainan, China, 2021, pp. 883–887, doi: 10.1109/QRS54544.2021.00097.
[31] S. Mahinnezhad, S. Mahinnezhad, K. Kaur, and A. Shih, "Data augmentation and class imbalance compensation using CTGAN to improve gas detection systems," in 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Glasgow, United Kingdom, 2024, pp. 1–6, doi: 10.1109/I2MTC60896.2024.10561121.
[32] M. A. I. Arif and M. M. Rahman, "SLDXG: CTGAN based semi-supervised learning model for network attack detection," in 2025 28th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2025, pp. 4861–4866, doi: 10.1109/ICCIT68739.2025.11491837.