

# Depression Detection in Bengali Social-Media Text via Clinically Grounded Linguistic Feature Fusion and Interpretable Machine Learning

**Md. Moinul Islam**

Daffodil International University <https://orcid.org/0009-0008-6588-4448>

**Md. Abdulla Al Noman**

Daffodil International University

**Most. Khadiza Akter Summa**

Daffodil International University

**S.K.M Shadikul Islam**

[shadikul15-12836@diu.edu.bd](mailto:shadikul15-12836@diu.edu.bd)

Daffodil International University

**Sarbajit Paul Bappy**

Daffodil International University

**Aminul Haque**

Daffodil International University

---

## Research Article

**Keywords:** depression detection, Bengali NLP, low-resource languages, psycholinguistic features, soft-voting ensemble, explainable AI, SHAP

**Posted Date:** October 27th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11493909/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Abstract

Depression ranks among the leading causes of disability worldwide, yet automated screening tools remain scarce for low-resource languages such as Bengali. In Bangladesh, where roughly one psychiatrist serves every 300,000 people and individuals are more likely to express distress in Bangla than in English, this gap carries real practical consequences. We present a transparent and fully reproducible pipeline for flagging depression-indicative posts in the BSMDD corpus, a balanced collection of 21,910 Bengali social-media posts. Rather than treating each post as an unordered bag of words, we encode five psycholinguistic markers of depressive cognition established in the clinical literature: negation, absolutist (all-or-nothing) language, hopelessness, self-focus, and temporal persistence, supplemented by a distress lexicon. Each marker is represented twice: as an injected lexical token that lets the n-gram model learn its surrounding context, and as a length-invariant numeric density. Mann-Whitney U tests confirm that all markers differ significantly between classes; distress, hopelessness, and self-focus densities run roughly three times higher in depression posts than in controls. We fuse word-level and character-level TF-IDF with this clinical block, calibrate three base learners (logistic regression, a linear SVM, and LightGBM), and combine them in a soft-voting ensemble whose weights and decision threshold are tuned solely on validation data. On a held-out test set examined exactly once, the ensemble achieves 91.66% accuracy and 91.69% F1, a ROC-AUC of 0.969, and a 95% bootstrap F1 interval of 90.68 to 92.62%. Global logistic regression coefficients, SHAP explain each prediction at both the global level. Manual review of misclassified posts indicates that most errors arise from noisy annotation rather than model failure. The outcome is a competitive, fully auditable, and low-cost baseline for Bengali mental-health NLP.

## 1. INTRODUCTION

Depression, formally major depressive disorder (MDD), is one of the leading causes of disability worldwide. Its hallmark features include persistent low mood, loss of interest in previously enjoyable activities, fatigue, impaired concentration, disrupted sleep, pervasive feelings of worthlessness, and, in serious cases, suicidal ideation. For a clinical diagnosis these symptoms must persist for at least two weeks and interfere meaningfully with daily functioning [1]. Because biological, psychological, and social factors all contribute, early and accurate identification remains both urgent and genuinely difficult [2].

The global burden is substantial. The World Health Organization estimates that roughly one in eight people lives with a mental disorder, with depression among the most prevalent, affecting approximately 280 million people or close to 3.8% of the global population [3]. That figure has climbed around 13% over the past decade, and the World Economic Forum places the associated annual productivity losses in the trillions of US dollars [2]. In 2022 alone, an estimated 49,369 people died by suicide linked to depressive disorders, and that toll continues to rise [4].

In Bangladesh, the situation is especially acute. Data from the Institute for Health Metrics and Evaluation covering 2010 to 2019 indicate that roughly 4.5 million Bangladeshis aged 20 to 54 experience

depression, with prevalence increasing approximately 1.9% each year [1]. Among university students, one study found a prevalence as high as

69.5%, up sharply from 24% in 2015 [5]. During the COVID-19 pandemic, approximately 3.7% of Bangladeshi college students with depression attempted suicide [5]. Stigma compounds the problem; those who seek help are often dismissed as attention-seekers [5]. Globally, around 70% of people avoid professional help in the early stages of depression [3], and in Bangladesh, where a single psychiatrist may serve approximately 300,000 people [5], this structural barrier is especially high.

Social media has opened a valuable channel for studying mental health at the population scale. Platforms such as Facebook, X (formerly Twitter), Instagram, and Reddit collectively generate billions of posts each day, and people routinely use them to express personal difficulty and shifting mental states [3]. Research shows that this language can reliably indicate depression and enables non-intrusive screening alongside conventional clinical methods [4]. Importantly, non-native English speakers tend to disclose distress more authentically in their own language [4], making Bangla social-media text a rich but still underexplored resource [6].

Machine learning and NLP have become the standard tools for depression detection in social-media text. Early pipelines relied on classical algorithms such as SVM, Naive Bayes, decision trees, and random forests applied to handcrafted surface features [7]. Deep learning architectures, including CNNs, RNNs, and LSTMs, later improved the modelling of sequential context [8]. Transformer models such as BERT, BanglaBERT, and XLM-RoBERTa pushed the field further

by capturing bidirectional long-range context [2], and recent large language models including GPT-3.5 and GPT-4 have approached state-of-the-art accuracy at considerable computational cost [1]. Despite this progress, the vast majority of work targets English; a post-COVID-19 review documented more than a 50% rise in global depression rates while noting that multilingual detection remains severely underexplored [9].

Bengali NLP has attracted growing interest in sentiment analysis [10] and low-resource language settings [11], but dedicated mental-health detection in Bangla remains rare. Cultural context shapes how depression surfaces in language, and predictive features that work in English do not transfer reliably across languages [4]. The few existing Bangla studies mostly rely on small or non-clinical datasets, rarely compare architectures systematically, and largely treat the task as a plain binary decision with little attention paid to clinical feature grounding, explainability, or deployment readiness [2, 12].

We present a transparent and reproducible framework grounded in clinical psychology for detecting depression-indicative posts in Bangla social-media text. The study has three concrete objectives. First, we operationalise five established psycholinguistic markers of depressive cognition in Bengali and confirm their discriminative power statistically. Second, we fuse heterogeneous text representations, benchmark seven classical learners under stratified cross-validation, and combine the strongest models in a calibrated soft-voting ensemble with proper significance testing. Third, we explain every prediction

globally through logistic-regression coefficients and SHAP, releasing a deployment-ready saved pipeline. The remainder of the paper covers related work in Section 2, methodology in Section 3, results and discussion in Section 4, and conclusions in Section 5.

## 2. RELATED WORK

Dabhane and Chawan [7] surveyed machine learning approaches to depression detection in social media and traced how classical pipelines built on handcrafted features initiated the field. Phiri et al. [3] subsequently conducted a systematic review and meta-analysis confirming that linguistic cues such as first-person pronouns, negative affect, and anxiety vocabulary are among the strongest predictors of depression in social-media text, although evidence for non-English settings remains limited. In the explainability domain, Hameed et al. [13] demonstrated that pipelines combining SVM, random forests, and gradient boosting with LIME improve clinical transparency; however, like most work in this space, theirs is confined to English. Within Bengali, Siddiqua et al. [14] achieved 91.1% accuracy with a random forest but on structured student survey data rather than free-form social-media posts. Padmaja et al. [8]

proposed a hybrid deep learning model combining NLP with BiLSTM architectures that handles sequential context effectively but has been evaluated primarily in English, and Bucur et al. [9] highlighted through a comprehensive post-pandemic review both the gains of such architectures and their weak cross-cultural generalisation.

Islam et al. [2] proposed a diversity-aware transformer ensemble with post-hoc explainability that ranks among the leading approaches for multi-label emotion detection and severity classification in English. The most directly relevant Bengali precedent is the work of Chowdhury et al. [1], who applied deep learning, transformer, and large language model architectures to binary depression detection in Bangla and found that fine-tuned LLMs performed best; however, their dataset was private and the work did not address severity grading or explainability. Within Bengali-specific research, Kabir et al. [12] constructed a DSM-5-aligned severity corpus on which a GRU reached 81% accuracy, while Siddik et al. [5] studied Bangladeshi university students and demonstrated the feasibility of ML-based early identification using classical classifiers. Further Bangla contributions include social-media depression detection [15], CNN-BiLSTM approaches for low-resource mental-health detection [16], hybrid CNN-BiGRU-BiLSTM classification [17], and a Bengali NLP prediction study published in PLOS ONE [18]. At the multilingual level, Bucur et al. [4] surveyed 108 datasets spanning 25 languages and explicitly identified Bangla as a language with significant unaddressed research gaps.

Table 1 summarises the studies reviewed above. Three recurring gaps emerge consistently across this literature. Bengali depression detection is substantially underexplored relative to English, as Bucur et al. [4] documented. The existing Bengali studies, such as that of Kabir et al. [12], rely on small corpora with limited clinically grounded feature design. And explainability, which Hameed et al. [13] have shown to be both feasible and clinically meaningful in English-language systems, is largely absent from Bangla

research. This work directly addresses all three gaps by combining a large balanced Bengali corpus, clinical feature design grounded in established psycholinguistic theory, and full post-hoc explainability.

Table 1  
Representative studies (Acc: accuracy; Rec: recall).

| Study                | Method                                 | Performance                     | Data               |
|----------------------|----------------------------------------|---------------------------------|--------------------|
| Siddiqua et al. [14] | RF, LR, KNN; BERT / RoBERTa; SHAP/LIME | RF 91.1% Acc; RoBERTa 98.6% Rec | Bengali (tabular)  |
| Kabir et al. [12]    | Kernel SVM, RF, LR, GRU / LSTM         | GRU 81% Acc;<br>SVM 78%         | Bengali (severity) |
| Chowdhur             | Deep learning,                         | LLMs best                       | Bengali            |

| Study                | Method                                                           | Performance                                     | Data                               |
|----------------------|------------------------------------------------------------------|-------------------------------------------------|------------------------------------|
| y et al. [1]         | transformers, LLMs                                               |                                                 | (private)                          |
| Islam et al. [2]     | Transformer ensemble + LIME                                      | Macro-F1 80.4%                                  | English (Reddit)                   |
| Padmaja et al. [8]   | NLP + hybrid deep learning                                       | State-of-art (Eng.)                             | English                            |
| Neog and Baruah [11] | RF, NB, SVM; LSTM, RNN                                           | LSTM 89.1%<br>Acc                               | Assamese                           |
| <b>This work</b>     | <b>Clinical feature fusion + soft-voting ensemble; SHAP/LIME</b> | <b>91.66% Acc;<br/>91.69% F1;<br/>AUC 0.969</b> | <b>Bengali (BSMDD, n = 21,897)</b> |

### 3. METHODOLOGY

The pipeline is designed for reproducibility and transparency throughout. Random seeds are fixed at 42, splits are deterministic, and every vectoriser and scaler is fit on the training partition only, ensuring nothing from the test set leaks back into training. All computations run on a CPU in Python 3.12 using scikit-learn 1.6, XGBoost, LightGBM, SHAP, and LIME. The pipeline operates in three broad stages: text cleaning and feature engineering; cross-validated model selection and ensemble construction; and held-out evaluation with statistical testing and post-hoc explanation. Figure 1 provides the complete architectural overview.

*(1) data acquisition and preparation from the BSMDD corpus; (2) Bengali text preprocessing; (3) clinical feature engineering and heterogeneous feature fusion; (4) model ablation study and proposed ensemble*

construction with validation-tuned weights and threshold; and (5) held-out evaluation with statistical testing and full post-hoc interpretability.

### 3.1 Dataset

The data come from the BSMDD corpus [20], in which each Bengali social-media post is labelled as depression-indicative (1) or control (0). There are 21,910 posts in total, with no missing values and six exact duplicates. The two classes are near-perfectly balanced: 10,949 control and 10,961 depression posts (49.97% and 50.03%, imbalance ratio 1.001), which means accuracy and F1 can be compared directly without any resampling. After removing the six

duplicates and a further seven posts that cleaning reduced to fewer than three tokens, 21,897 posts remained for modelling (Table 2). Crucially, post length offers no classifier shortcut: control posts average 81.1 words (median 60) and depression posts 80.1 words (median 63), nearly identical distributions confirmed in Fig. 2.

Table 2  
BSMDD dataset card after integrity audit and preprocessing.

| Property                 | Value            |
|--------------------------|------------------|
| Raw posts                | 21,910           |
| Missing values           | 0                |
| Exact duplicates removed | 6                |
| Control (label 0)        | 10,949 (49.97%)  |
| Depression (label 1)     | 10,961 (50.03%)  |
| Imbalance ratio          | 1.001 (balanced) |
| Posts after cleaning     | 21,897           |
| Total tokens             | 1,766,237        |
| Unique token types       | 61,238           |
| Type/token ratio         | 0.0347           |

### 3.2 Text Preprocessing

Bengali is a morphologically rich Indic language written in the Bengali-Assamese script (Unicode block U + 0980 to U+09FF). On social media it attracts code-mixing, transliteration noise, and inconsistent encodings. The cleaning pipeline normalises text to Unicode NFC, strips URLs, email addresses, mentions, Latin characters, digits, and non-Bengali symbols, collapses whitespace, and removes 80 common Bengali stop-words. One deliberate design choice stands out: negation markers and other

clinically meaningful words are retained rather than dropped, because terms such as **না** (not) and **হয়নি** (absent) carry diagnostic weight that would be lost during stop-word

removal. Posts shrinking to fewer than three tokens after cleaning are discarded as carrying insufficient affective signal. The full corpus cleans in about 4.4 seconds on a single CPU core.

### 3.3 Clinical Feature Engineering

The core novelty of the method is a small but targeted set of features that capture cues a clinician would notice but that a plain bag-of-words model systematically misses. We use five evidence-based markers of depressive cognition, namely negation, absolutist language, hopelessness, self-focus, and temporal persistence, plus a distress lexicon. Every marker is represented twice. It is first injected into the token stream: a negation is attached to the word it governs, so **না হয়নি** (cannot) becomes the token **NEG\_হয়নি**, giving the n-gram model explicit contextual information around each marker. It is then encoded as a length-normalised density: together with the log document length, these produce seven dense numeric features. Table 3 lists each lexicon with its clinical rationale and representative Bengali terms.

Table 3

Clinical lexicons with representative Bengali terms. Each lexicon contributes one length-normalised density feature, and its markers are also injected into the token stream.

| Marker (n)         | Clinical rationale                                                                                 | Bengali examples                              |
|--------------------|----------------------------------------------------------------------------------------------------|-----------------------------------------------|
| Negation (9)       | Reverses polarity; unscoped negation makes a positive word appear class-neutral.                   | না, হয়নি, নেই, নেইতো                         |
| Absolutis t (18)   | All-or-nothing thinking is elevated in depression and suicidal ideation [19].                      | সর্বসম্মত,<br>সর্বসম্মত, সর্বসম্মত, সর্বসম্মত |
| Hopeless ness (19) | Hopelessness is among the strongest predictors of suicidal risk (Beck Hopelessness Scale).         | সর্বসম্মত, সর্বসম্মত, সর্বসম্মত,<br>সর্বসম্মত |
| Self-focus (8)     | Heavy first-person singular use is a recognised marker of ruminative self-focus in depression [3]. | আমি, আমার, আমারই,<br>আমিই                     |
| Temporal (12)      | Depression is defined partly by chronicity: always sad differs clinically from sometimes sad.      | সর্বসম্মত, সর্বসম্মত,<br>সর্বসম্মত, সর্বসম্মত |
| Distress (20)      | Explicit affective-distress and self-harm vocabulary provides direct supervision-aligned signal.   | সর্বসম্মত, সর্বসম্মত, সর্বসম্মত,<br>সর্বসম্মত |

### 3.4 Feature Fusion and Modelling

Three complementary views of each post are constructed and concatenated into a single sparse matrix. Word-level TF-IDF uses unigrams and bigrams over marker-augmented text, with sublinear term-frequency scaling and a vocabulary cap of 25,000 features. Character n-gram TF-IDF covers 3- to 5-grams with the same feature budget, handling Bengali morphology, sandhi, and spelling variation in a way that word-level features cannot. The seven clinical features form the third view, scaled to the unit

range so the combined matrix remains non-negative and stays compatible with Naive Bayes. The fused training matrix has 50,007 dimensions and roughly 13.1 million non-zero entries.

Data are split with stratification into 70% training, 15% validation, and 15% test, yielding 15,326, 3,286, and 3,285 posts, respectively, each preserving the 50% class ratio. The test set is touched exactly once, for the final evaluation. Seven classifiers spanning the linear, probabilistic, kernel, tree-ensemble, and neural families are compared using stratified five-fold cross-validation. The three models with the highest cross-validated F1 that also complement each other, namely logistic regression, a calibrated linear SVM, and LightGBM, are combined in a weighted soft-voting ensemble whose voting weights and decision threshold are tuned on the validation set only. Predictions are then explained globally through logistic-regression coefficients and SHAP, and locally through LIME.

## 4. RESULTS AND DISCUSSION

**4.1 Discriminative Power of the Clinical Features** Each clinical marker reliably separates the two classes (Table 4; all p-values well below 0.001). Distress vocabulary carries the largest effect, with a rank-biserial coefficient of 0.435 and a mean density roughly 3.2 times higher in depression posts than in controls. Hopelessness and self-focus densities are also substantially elevated at 3.5x and 3.2x respectively, fully consistent with the clinical literature on rumination and hopelessness as core features of MDD. Temporal persistence and absolutist language show more modest but still significant elevation, while negation, the weakest individual marker, nevertheless reaches significance across more than 21,000 observations. Figure 3 visualises these class-level differences across all six markers. These are not arbitrary text statistics; they are the exact cues clinicians already watch for, and it is meaningful to confirm they surface reliably in Bengali social-media text.

Table 4  
Mann-Whitney U validation of the clinical features, sorted by absolute rank-biserial effect size.

| Feature                 | Mean (dep) | Mean (ctrl) | Ratio | p-value  | rbc          |
|-------------------------|------------|-------------|-------|----------|--------------|
| <i>distress_density</i> | 0.0291     | 0.0092      | 3.15  | < 1e-300 | <b>0.435</b> |

| Feature             | Mean (dep) | Mean (ctrl) | Ratio | p-value   | rbc   |
|---------------------|------------|-------------|-------|-----------|-------|
| <i>hope_density</i> | 0.0133     | 0.0038      | 3.5   | < 1e-300  | 0.292 |
| <i>self_density</i> | 0.0097     | 0.003       | 3.24  | < 1e-300  | 0.251 |
| <i>log_len</i>      | 4.072      | 4.227       | 0.96  | 5.20E-80  | 0.148 |
| <i>temp_density</i> | 0.0092     | 0.0055      | 1.69  | 1.60E-101 | 0.141 |
| <i>abs_density</i>  | 0.0173     | 0.0137      | 1.26  | 3.40E-28  | 0.082 |
| <i>neg_density</i>  | 0.00068    | 0.00039     | 1.77  | 8.60E-16  | 0.021 |

| Model         | Acc%      | Prec%  | Rec%   | F1%    | AUC<br>% |
|---------------|-----------|--------|--------|--------|----------|
| Logistic Reg. | 88.95 ± 0 | 88.72± | 89.28± | 89.00± | 95.47±   |
|               | .30       | 0.61   | 0.53   | 0.28   | 0.27     |
| MLP           | 88.44 ± 0 | 87.20± | 90.16± | 88.64± | 95.10±   |
|               | .38       | 0.94   | 1.43   | 0.43   | 0.25     |
| Random Forest | 87.79 ± 0 | 86.02± | 90.27± | 88.09± | 94.71±   |
|               | .62       | 0.82   | 0.60   | 0.58   | 0.25     |
| Complement NB | 87.51 ± 0 | 85.01± | 91.11± | 87.95± | 94.23±   |
|               | .45       | 0.25   | 0.98   | 0.49   | 0.26     |

## 4.2 Model Ablation and Feature Fusion Study

In five-fold cross-validation on the training set (Table 5), LightGBM achieves the best single-model F1 of 90.80% and ROC-AUC of 96.66%, with XGBoost close behind at 90.30% F1. The calibrated linear SVM and logistic regression both sit near 89% F1, which is remarkably competitive with the tree models and suggests the fused feature space is nearly linearly separable. Tree models add roughly one to two F1 points by capturing non-linear interactions among the clinical densities, though at substantially higher training cost. Standard deviations remain at or below 0.6% across all folds, confirming that the performance ranking is stable and not an artefact of a single split. Figure 4 visualises the F1 scores and fold-level variance for all seven classifiers.

Table 5  
Stratified 5-fold CV on training set (mean  $\pm$  std%). Best model in bold.

| Model           | Acc%                            | Prec%                        | Rec%                         | F1%                          | AUC<br>%                     |
|-----------------|---------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| <b>LightGBM</b> | <b>90.72 <math>\pm</math> 0</b> | <b>90.07<math>\pm</math></b> | <b>91.54<math>\pm</math></b> | <b>90.80<math>\pm</math></b> | <b>96.66<math>\pm</math></b> |
|                 | <b>.39</b>                      | <b>0.44</b>                  | <b>0.94</b>                  | <b>0.42</b>                  | <b>0.19</b>                  |
| XGBoost         | 90.24 $\pm$ 0                   | 89.77 $\pm$                  | 90.85 $\pm$                  | 90.30 $\pm$                  | 96.40 $\pm$                  |
|                 | .44                             | 0.42                         | 0.98                         | 0.47                         | 0.22                         |
| Linear SVM      | 89.16 $\pm$ 0                   | 88.83 $\pm$                  | 89.61 $\pm$                  | 89.21 $\pm$                  | 95.51 $\pm$                  |
|                 | .36                             | 0.80                         | 0.46                         | 0.32                         | 0.25                         |

The feature-fusion ablation in Table 6 demonstrates why all three representation views are retained. Word-level features alone reach 85.77% validation F1. Adding character n-grams delivers the single largest gain, approximately 2.3 percentage points, reflecting the importance of sub-word information for a morphologically complex language such as Bengali. The seven clinical features add a further measurable increment on top of the full lexical combination, and the complete fusion reaches 89.10% F1. Seven dimensions against fifty thousand may seem negligible, but the clinical block carries signal that the lexical features do not capture on their own, and that signal corresponds precisely to the markers clinicians find diagnostically meaningful.

Table 6  
Feature-fusion ablation with logistic regression on the validation set.

| Representation                       | Dims          | Val Acc%     | Val F1%     |
|--------------------------------------|---------------|--------------|-------------|
| Word TF-IDF only                     | 25,000        | 85.82        | 85.77       |
| Word + Character TF-IDF              | 50,000        | 88.07        | 88.08       |
| Word + Clinical block                | 25,007        | 86.03        | 86.1        |
| <b>Word + Char + Clinical (full)</b> | <b>50,007</b> | <b>89.07</b> | <b>89.1</b> |

## 4.3 Ensemble Performance on Held-Out Data

On the full training split, the three base learners scored validation F1 of 0.8910 (logistic regression), 0.8914 (calibrated SVM), and 0.9012 (LightGBM). Grid search over the validation set assigned voting weights of 1, 2, and 2 to the three models respectively, and the default decision threshold of 0.50 proved optimal, raising the validation F1 to 0.9091. On the held-out test set, examined exactly once as intended

from the outset, the ensemble achieved 91.66% accuracy, 91.41% precision, 91.97% recall, 91.69% F1, a ROC-AUC of 0.9687, and a PR-AUC of 0.9625 (Table 7).

The two classes score almost identically across every metric, which is reassuring given the near-perfect class balance. Recall slightly exceeds precision, which is exactly the right bias for a screening tool where missing a case carries a higher cost than raising a false alarm.

The 95% bootstrap confidence intervals are reassuringly tight: accuracy 90.68 to 92.51%, F1 90.68 to 92.62%, and ROC-AUC 96.28 to 97.36%. The calibration curve in Fig. 5(d) stays close to the perfect-calibration diagonal across the full probability range, with a Brier score of 0.065, confirming that the output probabilities are trustworthy for threshold-based deployment. Clinicians or system designers can shift the decision threshold toward higher recall in settings where missing a case is considered unacceptable.

Table 7  
Held-out test performance of the proposed ensemble at threshold = 0.50.

| Metric      | Overall         | Control | Depressi on | Macro  |
|-------------|-----------------|---------|-------------|--------|
| Accuracy    | 91.66%          | N/A     | N/A         | 0.9166 |
| Precision   | 91.41%          | 0.9191  | 0.9141      | 0.9166 |
| Recall      | 91.97%          | 0.9135  | 0.9197      | 0.9166 |
| F1-score    | 91.69%          | 0.9163  | 0.9169      | 0.9166 |
| ROC-AUC     | 0.9687          | N/A     | N/A         | N/A    |
| PR-AUC      | 0.9625          | N/A     | N/A         | N/A    |
| 95% CI (F1) | [90.68, 92.62]% | N/A     | N/A         | N/A    |
| Support     | 3,285           | 1,641   | 1,644       | N/A    |

The ensemble does not dramatically outperform its best individual component in raw accuracy. In McNemar's test against LightGBM alone, 2,940 of 3,285 test posts were correctly classified by both models, 63 only by LightGBM, 71 only by the ensemble, and 211 by neither, yielding a chi-squared statistic of 0.366 and p of 0.545. On this balanced corpus the ensemble statistically ties LightGBM rather than clearly surpassing it. However, the ensemble's main advantages lie elsewhere: better probability calibration across the full score range and lower variance from combining three learners that exploit different aspects of the feature space. Both qualities matter more than raw accuracy in a threshold-based clinical screening setting.

## 4.4 Interpretability and Error Analysis

The logistic-regression coefficients in Fig. 6 place injected clinical markers, particularly hopelessness and self-focus tokens, alongside explicit distress vocabulary at the top of the depression-indicative feature list. This is precisely what we hoped to see: the model is learning genuine clinically grounded signal rather than spurious correlations. The reasoning behind any positive prediction is therefore readable and clinically plausible. On the control side, the top features are dominated by character n-grams and the log-length feature, suggesting that control posts tend to be longer and contain more varied sub-word patterns.

A LightGBM trained solely on the seven clinical features achieves 79.75% test F1, a surprisingly strong result for a seven-feature model and clear evidence that the clinical block carries real predictive weight. Its SHAP summary in Fig. 7 shows that higher distress, hopelessness, and self-focus densities shift predictions toward depression (red points to the right of centre), while longer documents push toward the control class. The negative SHAP contribution of log\_len for depression is noteworthy: depression posts in this corpus tend to be slightly shorter on average, perhaps because distress is expressed urgently and tersely rather than expansively.

The ensemble misclassified 274 of the 3,285 test posts (8.34%), with 142 false positives and 132 false negatives. Inspecting the most confidently misclassified false positives reveals something instructive: several control-labelled posts openly discuss suicidal ideation and self-harm in language that is linguistically indistinguishable from depression posts. The model is not failing on these cases; the annotations appear to be incorrect. This suggests that the headline test figures likely understate the model's true discriminative capability, with some apparent mistakes reflecting mislabelled ground truth rather than genuine model errors.

Taking a step back, this work confirms that depressive cognition leaves measurable and consistent traces in Bengali text, much as it does in English. A small, interpretable set of clinical features captures a substantial fraction of that signal without sacrificing transparency. Character-level modelling delivers the largest single representation gain, which is expected for a morphologically rich language with high spelling variation. The full pipeline reaches 91.7% F1 on modest hardware while remaining fully explainable, making it a practical and trustworthy baseline in a context where clinical transparency and reproducibility matter as much as raw accuracy [1].

## 4.5 Ethical Considerations and Limitations

This is a screening aid, not a diagnostic tool, and it should never be used to label an individual without clinician oversight and informed consent. Because missing a case costs more than a false alarm, the decision threshold should be tuned toward higher recall in any real deployment, and anyone the system flags should be routed to a human clinician with crisis resources available. Social-media text is sensitive personal data and requires appropriate ethics approval and platform compliance before collection or use.

Several limitations are clear. The clinical lexicons are hand-built and conservative; a validated, morphologically-aware Bengali affective lexicon would substantially improve marker recall. Label quality sets an upper bound on performance, and the error analysis confirms that some labels are genuinely ambiguous. The model reflects a single corpus from a specific time period, so generalisation to other platforms, dialects, or eras should be tested before deployment. Finally, a social-media post captures expressed affect, which is only a proxy for a clinical state: the model predicts how a post reads, not what its author is experiencing.

## 5. CONCLUSION AND FUTURE WORK

We described a transparent and reproducible pipeline grounded in clinical psychology for detecting depression-indicative posts in Bengali social-media text. Encoding five psycholinguistic markers and a distress lexicon as both lexical tokens and numeric densities, fusing them with word-level and character-level TF-IDF, and combining three calibrated learners in a validation-tuned soft-voting ensemble yields 91.66% accuracy and 91.69% F1 on a held-out test set, with well-calibrated output probabilities and explanations available at both the global and local level. Statistical significance testing and close inspection of the errors separate genuine model signal from annotation noise. The result is an auditable, low-cost baseline for a clinical problem that is both urgent and underserved.

The most immediate extension is running BanglaBERT and large language models on the identical data splits, which

would quantify the accuracy-versus-computational-cost trade-off that practitioners must navigate. The binary framing could grow into DSM-5-aligned severity grading [12], providing finer-grained information to clinicians. A Bengali morphological analyser and expanded lexicons reviewed by mental-health professionals would improve marker coverage substantially. Subgroup fairness audits and a prospective, ethics-approved clinical validation study are prerequisites before any real-world deployment. Releasing the corpus and the saved pipeline publicly would allow the research community to reproduce, extend, and challenge this work directly.

## Declarations

### CRedit Authorship Contribution Statement

**Md. Moinul Islam:** Conceptualization, Methodology, Data curation, Writing (original draft). **Md. Abdulla Al Noman and Most. Khadiza Akter Summa:** Data curation, Formal analysis, **S.K.M Shadikul Islam:** Project Administration, Writing (review and editing), **Sarbajit Paul Bappy:** Visualization, Validation, Writing (review and editing), **Aminul Haque:** Supervision, Writing (review and editing)

### Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

## References

1. Chowdhury AK, Sujon SR, Shafi MSS, Ahmmad T, Ahmed S, Hasib KM, Shah FM (2024) Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study. *Nat Lang Process J* 7:100075
2. Islam S, Haque R, Khan MA, Mohiuddin AB, Siddiqui MIH, Limon ZH, Appaji A (2026) Ensemble transformer with post-hoc explanations for depression emotion and severity detection. *iScience*, 29(2)
3. Phiri D, Makowa F, Amelia VL, Phiri YVA, Dlamini LP, Chung MH (2025) Text-based depression prediction on social media using machine learning: Systematic review and meta-analysis. *J Med Internet Res* 27:e59002
4. Bucur AM, Zampieri M, Ranasinghe T, Crestani F (2026), March A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 893–918)
5. Siddik MAB, Ali A, Miah S, Hasan M, Ahmed M, Sunna TC (2024) Psychological disorders among college going students: A post Covid-19 insight from Bangladesh. *J Affect Disorders Rep* 15:100686
6. Kabir MK, Islam M, Kabir ANB, Haque A, Rhaman MK (2022) Detection of depression severity using Bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Res*, 6(9), e36118
7. Dabhane S, Chawan PM (2020) Depression detection on social media using machine learning techniques: A survey. *Int Res J Eng Technol* 7(11):97–100
8. Padmaja SM, Godla SR, Ramesh JVN, Muniyandy E, Sridevi P, El-Ebiary YAB, Devadhas DN P (2025) Depression detection in social media using NLP and hybrid deep learning models. *Int J Adv Comput Sci Appl*, 16(2)
9. Bucur AM, Moldovan AC, Parvatikar K, Zampieri M, KhudaBukhsh AR, Dinu LP (2025) On the state of NLP approaches to modeling depression in social media: A post-COVID-19 outlook. *IEEE Journal of Biomedical and Health Informatics*
10. Ullah MA (2025) BCLSA: Advancing Bangla Sentiment Analysis with Concept-Level Reasoning and Efficiency. *Machine Learning with Applications*, p 100793
11. Neog M, Baruah N (2024) A hybrid deep learning approach for Assamese toxic comment detection in social media. *Procedia Comput Sci* 235:2297–2306
12. Kabir MK, Islam M, Kabir ANB, Haque A, Rhaman MK (2022) Detection of depression severity using Bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Res*, 6(9), e36118

13. Hameed S, Nauman M, Akhtar N, Fayyaz MA, Nawaz R (2025) Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. *Front Artif Intell* 8:1627078
14. Siddiqua R, Islam N, Bolaka JF, Khan R, Momen S (2023) AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students. *Array* 18:100291
15. Shwetha CH, Pushpalatha K (2026) A hybrid deep learning framework for multilingual depression detection and symptom classification from social media text. *Int J Inform Technol*, 1–24
16. Absar N, Islam MM, Somaya ZN (2025) Explainable depression detection from low-resource languages using CNN-BiLSTM with deep attention mechanism. *Mach Learn Comput Sci Eng* 1(2):29
17. Alhashmi SM, Hashim S, Khedr AM (2025), February Depression detection using a hybrid CNN-BiGRU-BiLSTM model with BERT embedding and data augmentation. In 2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 340–347). IEEE
18. Ayman U, Hasan MN, Khan MN, Saha MC, Fayejullah M (2025) Navigating tenses in Bengali sentences: A stacked ensemble model for enhanced prediction. *PLoS ONE*, 20(8), e0330186
19. Al-Mosaiwi M, Johnstone T (2018) In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clin Psychol Sci* 6(4):529–542
20. Chowdhury AK, Sujon SR, Shafi MSS, Ahmmad T, Ahmed S, Hasib KM, Shah FM (2024) Bengali Social Media Depressive Dataset (BSMDD). Mendeley Data, V1. <https://doi.org/10.17632/pdtyb7df6z.1>

## Figures

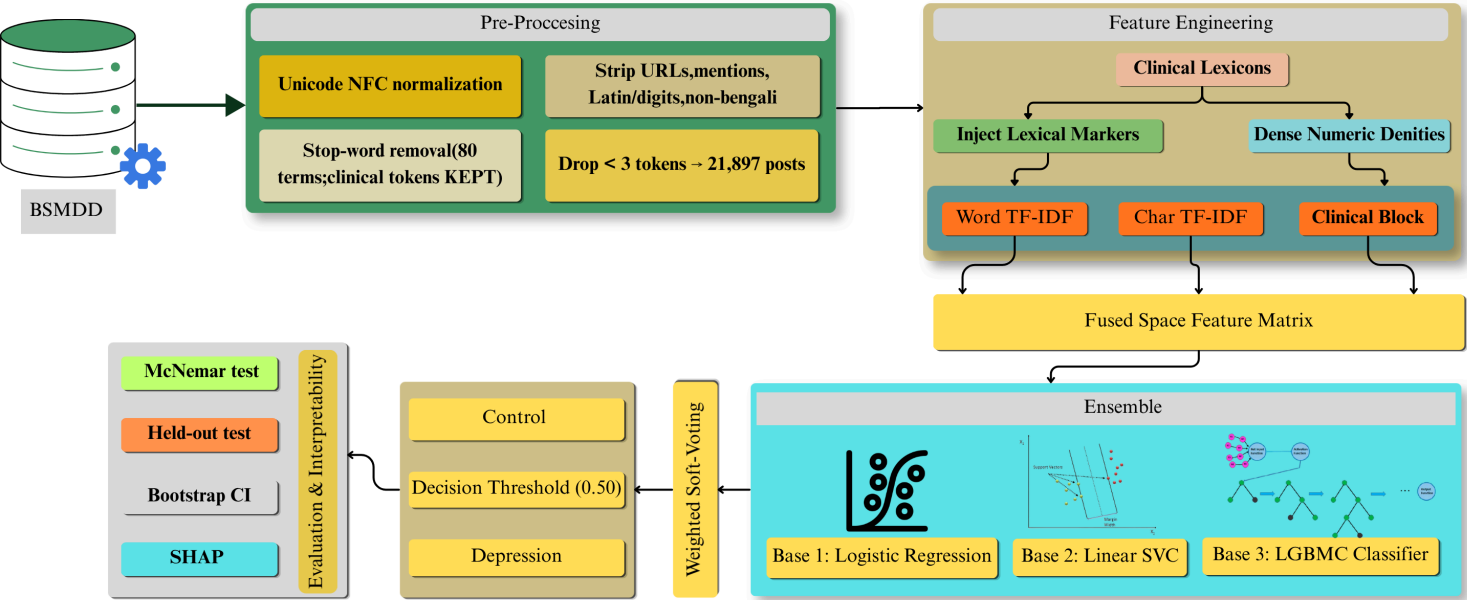


Figure 1

Methodology architecture. The pipeline has five stages:

(1) data acquisition and preparation from the BSMDD corpus; (2) Bengali text preprocessing; (3) clinical feature engineering and heterogeneous feature fusion; (4) model ablation study and proposed ensemble construction with validation-tuned weights and threshold; and (5) held-out evaluation with statistical testing and full post-hoc interpretability.

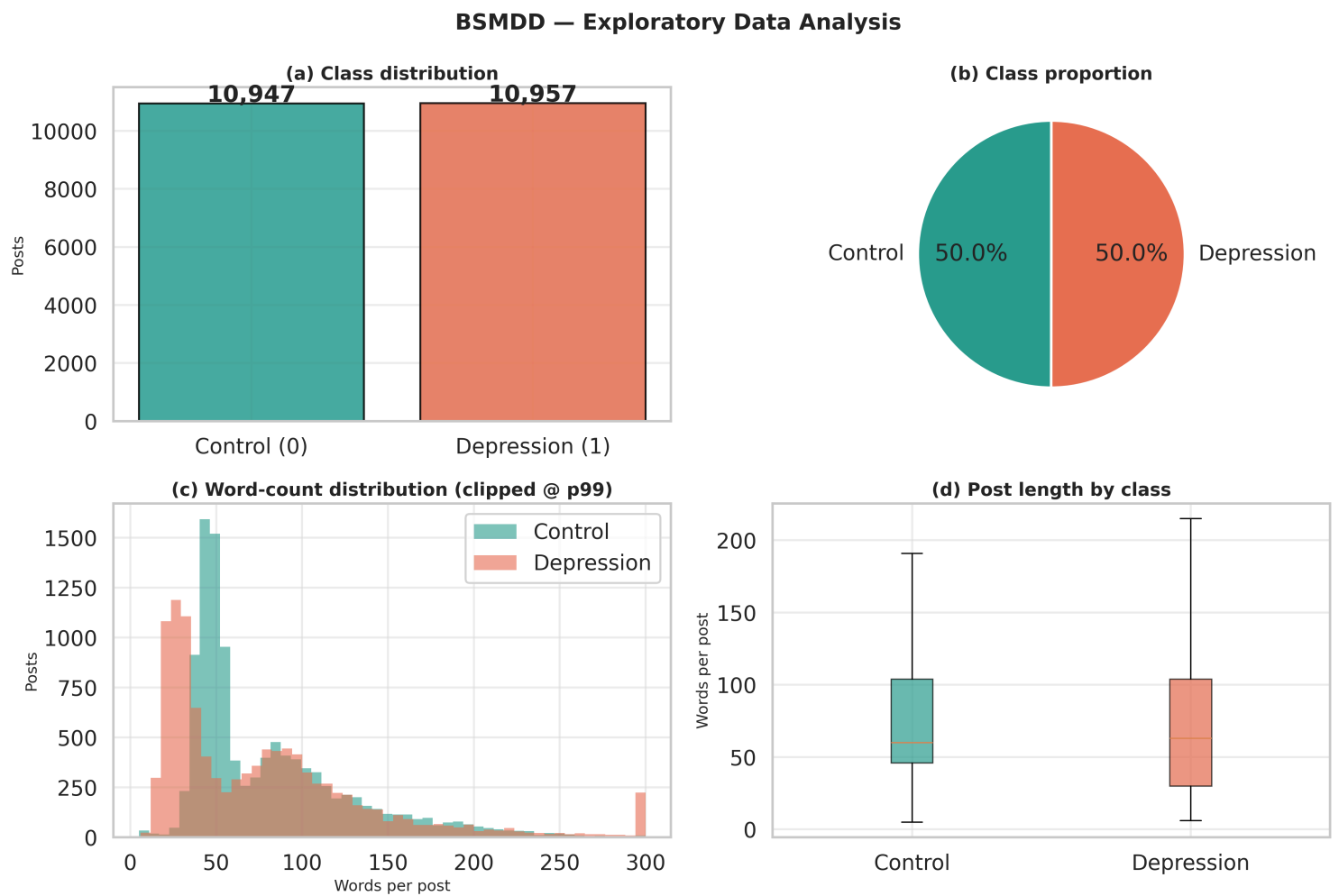


**Figure 2**

Exploratory analysis of BSMDD. (a) Class counts; (b) class proportion; (c) word-count distributions clipped at the 99th percentile; (d) post-length box plots. The two classes overlap heavily in length, ruling out length as a classifier shortcut.

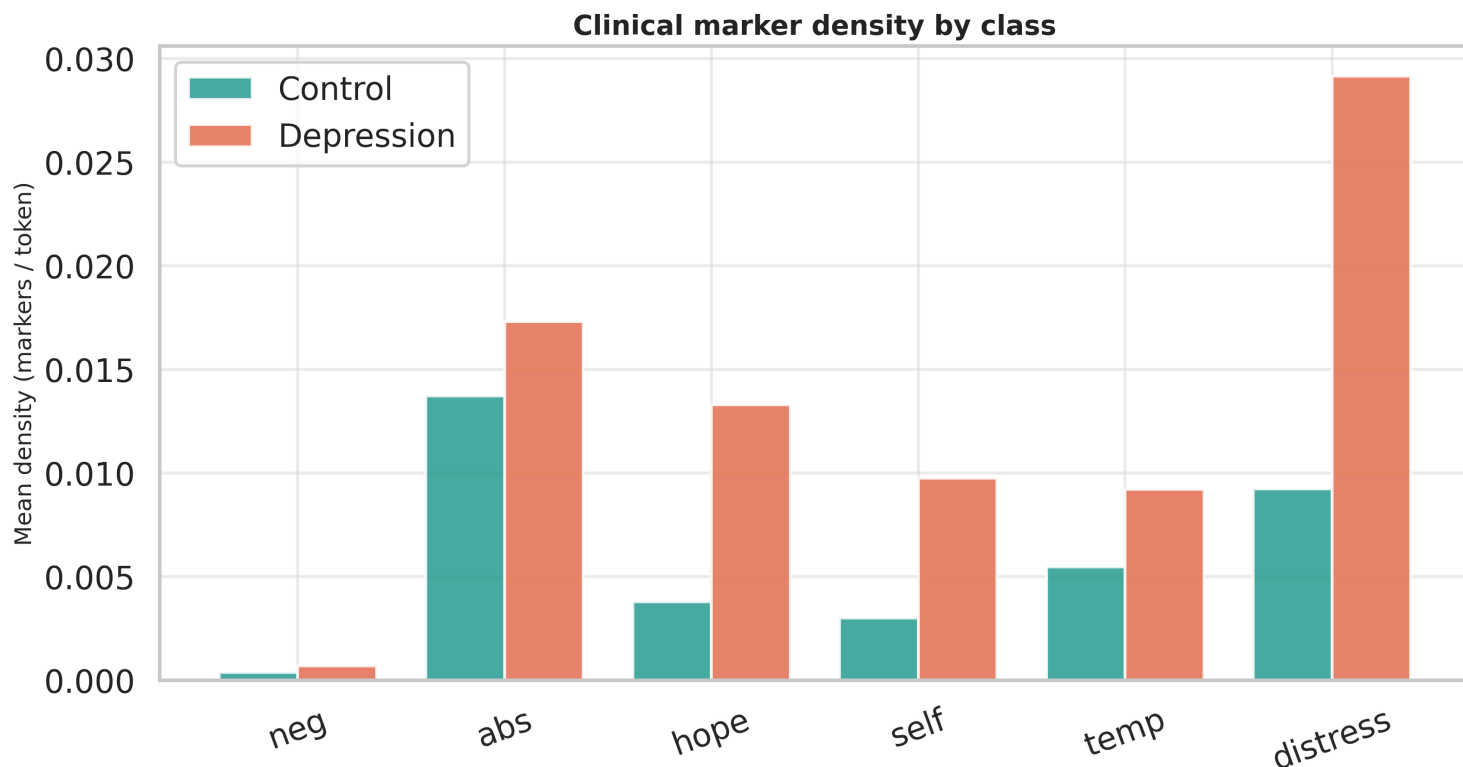


**Figure 3**

Mean clinical-marker density by class. Depression posts show substantially higher densities across all six markers, with distress vocabulary exhibiting the largest absolute gap at roughly 3 times the control level.

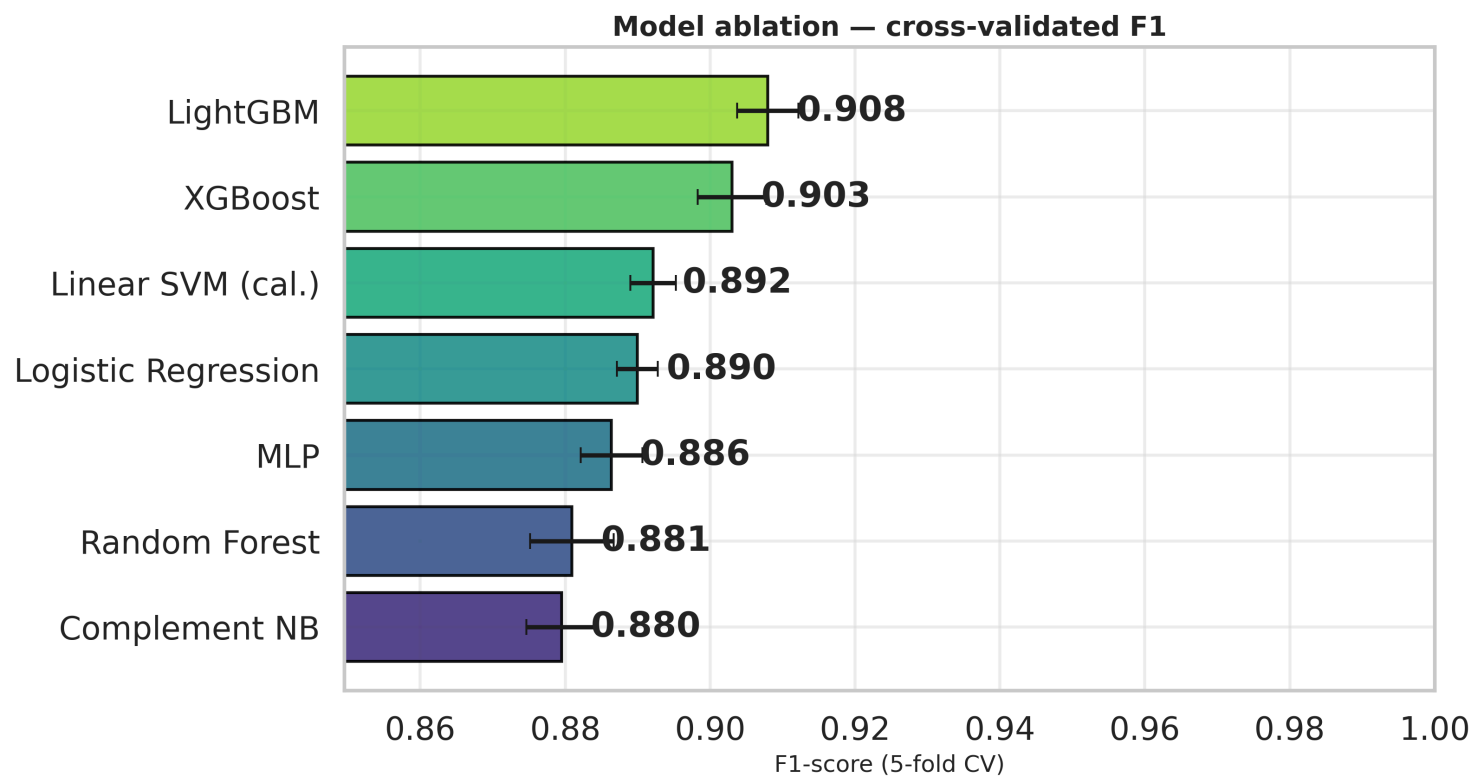


**Figure 4**

Model ablation: cross-validated F1 scores for all seven classifiers across five folds. LightGBM leads at 0.908 with XGBoost close behind at 0.903. Error bars indicate one standard deviation across folds.

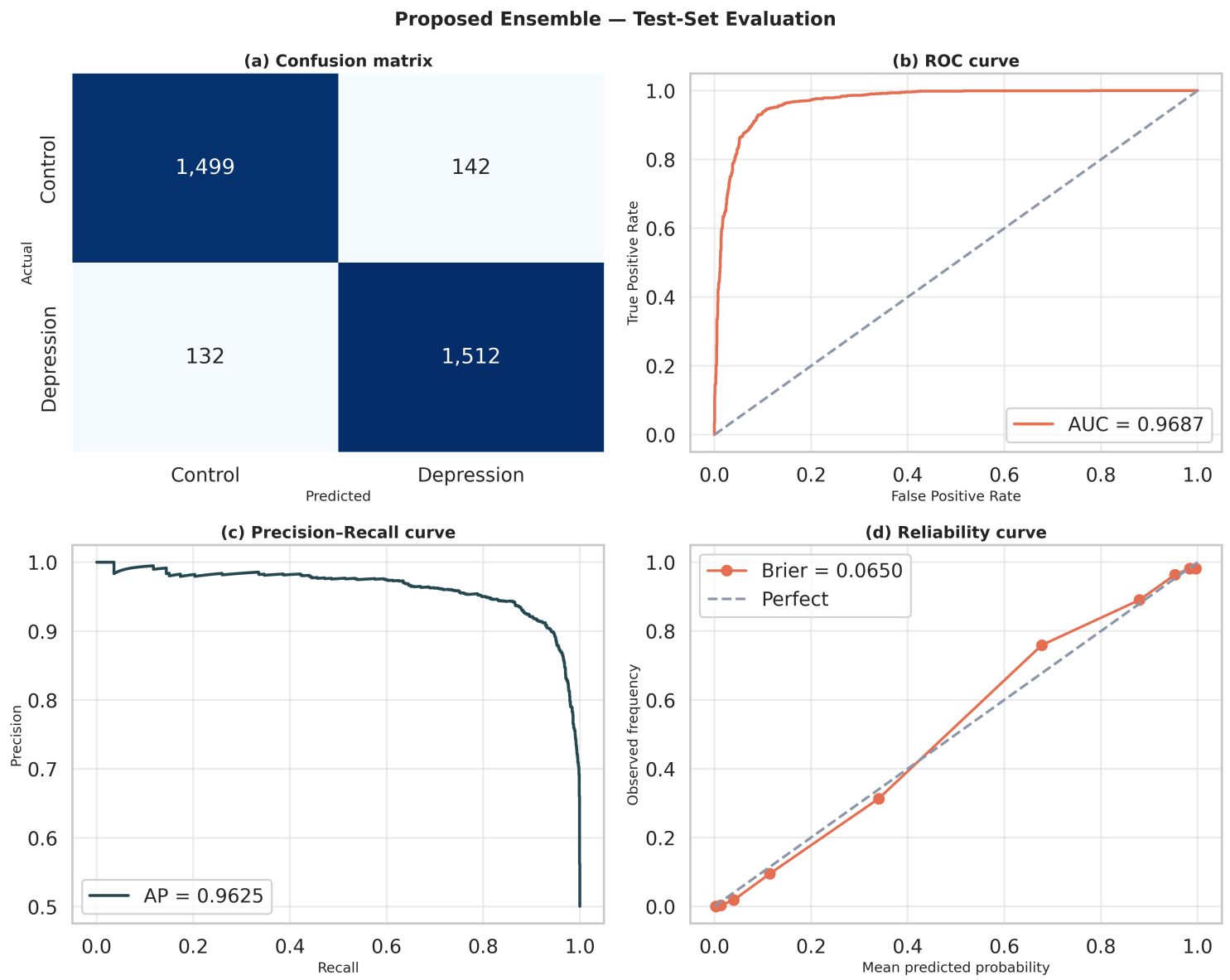


**Figure 5**

Test-set evaluation panel. (a) Confusion matrix: 1,499 true controls, 1,512 true depression cases, 142 false positives, 132 false negatives. (b) ROC curve (AUC 0.969). (c) Precision-recall curve (AP 0.963). (d) Reliability (calibration) curve with Brier score 0.065, staying close to the perfect-calibration diagonal throughout.

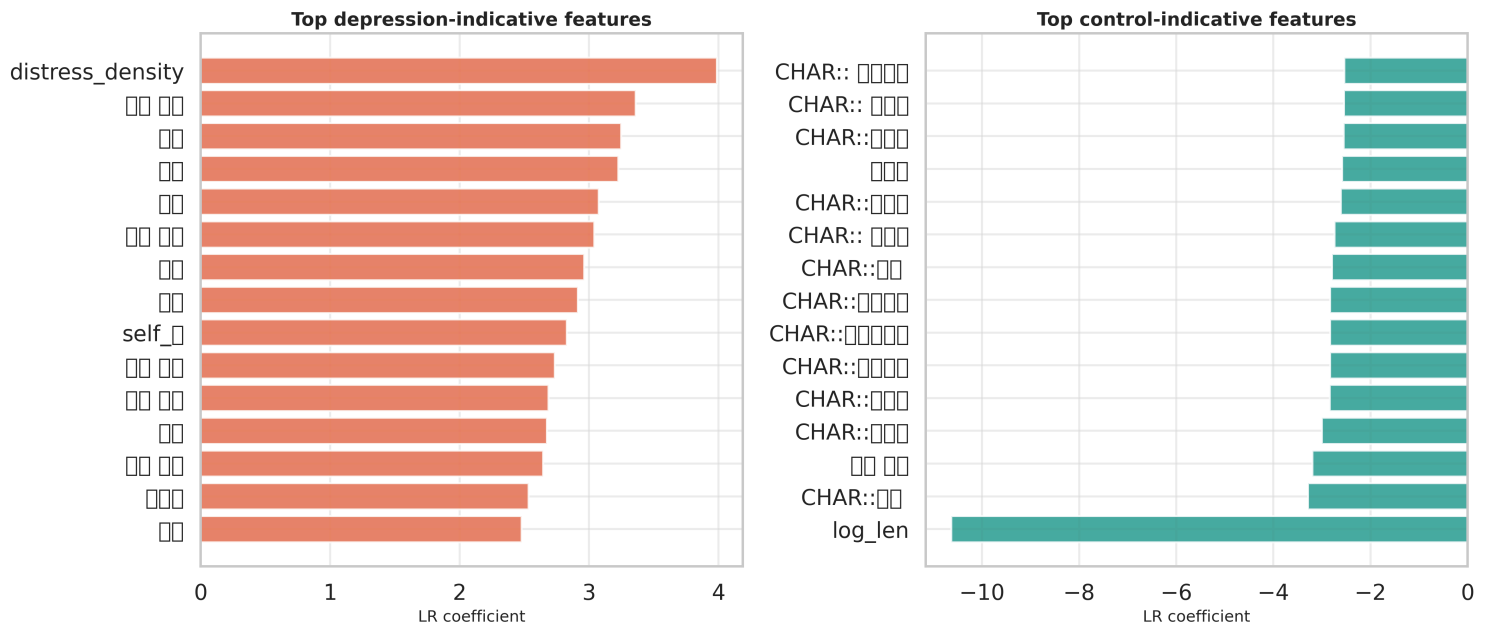


Figure 6

Top logistic-regression coefficients. Left panel: the 15 strongest depression-indicative features, led by `distress_density` and injected hopelessness / self-focus tokens. Right panel: the 15 strongest control-indicative features, dominated by character *n*-grams and `log_len`.

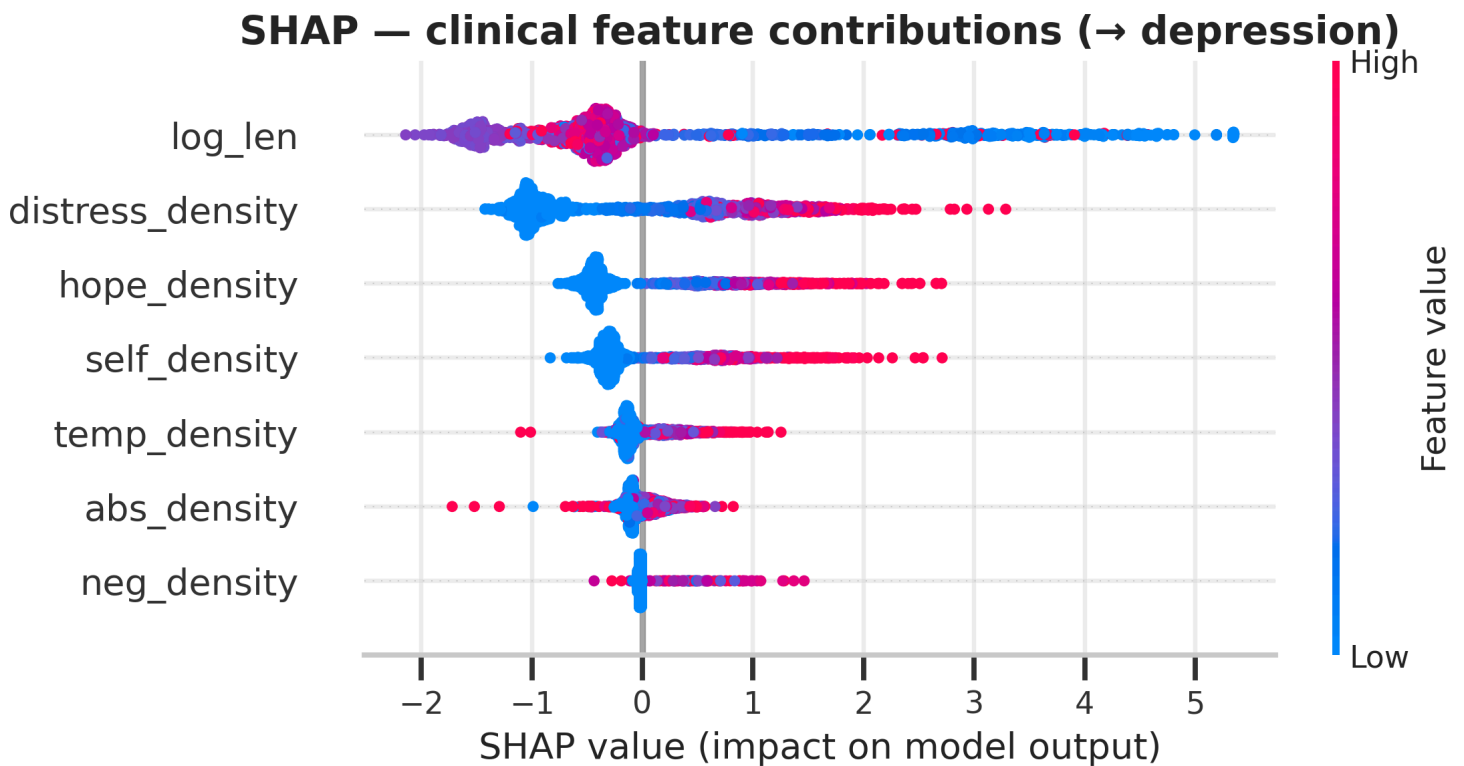


Figure 7

*SHAP summary for a clinical-only LightGBM (7 features, 2,000 test samples). High distress, hopelessness, and self-focus densities (red points, right of centre) increase the predicted probability of depression. High log-length consistently pushes toward the control class.*