ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

Early Stage Depression Recognition Using Machine Learning and Natural Language Processing: A Hybrid PHQ-9 and Linguistic Feature-Based Screening Framework

Tanu Chauhan[1]; Abhishek Verma[2]

[1,2]Department of MCA, Raj Kumar Goel Institute of Engineering and Technology, Ghaziabad

Publication Date: 2026/08/31

Abstract: Depression is a widespread and frequently under-diagnosed mental health condition, and delays in identification are associated with poorer long-term outcomes. This paper presents a hybrid screening framework that combines a validated questionnaire to the Patient Health Questionnaire-9 (PHQ-9), with linguistic feature extraction from short free-text journal entries to generate a composite depression-risk score. Unlike approaches that rely on a single data modality, the proposed system fuses structured clinical scoring with textual indicators associated with depressive symptomatology in prior computational-linguistics research, such as elevated first-person pronoun usage, negation patterns, and lexical markers of sadness, anhedonia, and worthlessness. A prototype web application implementing this framework was developed, comprising a screening interface, a relational database schema for longitudinal storage, and an analytics dashboard for aggregate monitoring. This paper describes the system architecture, the feature-fusion methodology, the underlying database design, and a discussion of how the prototype's rule-based linguistic module can be replaced with a trained supervised classifier (e.g., Support Vector Machine, Random Forest, or a fine-tuned transformer model) in future work. The framework is intended as a low-cost, scalable screening aid to support — not replace — clinical [1] evaluation.

Keywords: *Depression Recognition; Machine Learning; Natural Language Processing; PHQ-9; Mental Health Screening; Text Classification; Linguistic Analysis.*

How to Cite: Tanu Chauhan; Abhishek Verma (2026) Early Stage Depression Recognition Using Machine Learning and Natural Language Processing: A Hybrid PHQ-9 and Linguistic Feature-Based Screening Framework. *International Journal of Innovative Science and Research Technology*, 11(8), 2190-2197. https://doi.org/10.38124/ijisrt/26aug324

I. INTRODUCTION

Depression is one of the most leading causes of disability , yet a substantial proportion of affected individuals are never formally screened or diagnosed, largely due to stigma, limited access to mental health professionals, and the time-intensive nature of clinical assessment. Standardized self-report instruments such as the PHQ-9 remain the most widely used first-line screening tool in primary care, but they depend on active, voluntary disclosure and are typically administered only during a clinical encounter.

In parallel, the growth of machine learning (ML) and natural language processing (NLP) has made researchers to know the language people use — in social media posts, journal entries, and clinical notes — to identify linguistic markers associated with depressive states, including increased self-focus, negative affect words, and reduced social references. This has motivated a growing body of work on automated or semi-automated depression screening that supplements structured questionnaires with unstructured text analysis.

This paper proposes and implements a hybrid screening framework that fuses PHQ-9 responses with NLP-derived linguistic features extracted from a short free-text entry, producing a single composite risk score together with an interpretable breakdown of contributing factors. The contribution of this work is threefold: (i) a system architecture and relational database design for hybrid clinical-linguistic screening; (ii) a transparent, rule-based linguistic feature extraction module suitable as a baseline or interpretability layer alongside trained ML classifiers; and (iii) a working prototype with a participant-facing screening interface and a researcher-facing analytics dashboard.

II. RELATED WORK

Early work on social-media-based depression detection demonstrated that behavioral and linguistic signals extracted from platforms such as Twitter could distinguish users who

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

later self-reported a clinical depression diagnosis from control users, using features relating to social engagement, emotional expression, and linguistic style, including elevated first-person pronoun usage [1]. Building on this line of work, subsequent studies combined the PHQ-9 instrument directly with social media text: Motade et al. trained a Support Vector Machine classifier on a PHQ-9-labeled dataset alongside a Twitter dataset, reporting substantially higher classification accuracy on the structured PHQ-9 data than on informally written social media text, illustrating the value of combining structured and unstructured sources [2].

A recent systematic review and meta-analysis of language-based depression detection surveyed the feature-extraction and classification methods used across the field, noting that most studies define ground-truth depression status using a clinical research on a validated clinical questionnaire such as the PHQ-9, which supports the design choice in this paper of anchoring the composite score to PHQ-9 severity bands [5]. A related experimental case study specifically compared multiple ML classifiers and preprocessing pipelines for depression detection, highlighting data cleaning, feature selection, and model choice as key sources of performance variation between studies [4]. Outside of text-only approaches, researchers have also explored passively collected smartphone sensor data as a complementary, non-intrusive signal, training supervised classifiers to predict PHQ-9-derived depression severity from behavioral markers such as usage patterns and movement, reinforcing the broader trend toward multi-modal depression screening [3]. Most recently, large language models have been investigated directly for PHQ-9-style symptom labeling and reasoning over user-generated text, though this line of work also underscores open concerns about the reliability and quality control of machine-generated clinical labels [6].

The present work differs from these prior systems in scope: rather than proposing a new trained classifier, it contributes a full-stack reference architecture — screening interface, feature-fusion logic, relational schema, and analytics layer — intended to make the hybrid PHQ-9 + NLP approach reproducible and extensible, with an explicit, swappable interface between the transparent baseline linguistic module used in the prototype and a trained supervised or transformer-based classifier for future evaluation.

III. PROPOSED SYSTEM AND METHODOLOGY

➢ *System Architecture*

The proposed system follows a five-stage pipeline: (1) data collection via a structured PHQ-9 questionnaire and a free-text journal entry; (2) text preprocessing, including tokenization, stopword handling, and negation-aware matching; (3) feature extraction, comprising the PHQ-9 total score and a set of NLP-derived linguistic indicators; (4) classification, in which the two feature streams are fused into a composite risk score and mapped to a severity band; and (5) storage and analytics, in which each screening event is persisted for longitudinal and cohort-level analysis.

➢ *Clinical Feature Stream (PHQ-9)*

The PHQ-9 consists of nine questions, each points 0 (not at all) to 3 (nearly every day), yielding a raw total in the range 0-27. This raw score is linearly rescaled to a 0-100 range to place it on a common scale with the linguistic risk score described below. Item 9, which concerns thoughts of self-harm, is additionally tracked as a standalone binary flag, consistent with standard clinical practice of treating this item as a priority indicator regardless of overall score.

➢ *Linguistic Feature Stream (NLP)*

The free-text entry is processed through a lexicon-based feature extractor that counts occurrences of terms associated, in prior computational-linguistics research on depression, with six categories: sadness/dysphoria, fatigue, anhedonia, worthlessness, self-harm language, and social isolation. In addition, the density of first-person singular pronouns is computed, motivated by findings that increased self-focus is associated with depressive symptomatology. Category counts are normalized by entry length and combined into a single NLP risk score in the range 0-100. This module is intentionally implemented as a transparent, rule-based baseline for the prototype; Section 6 discusses its replacement with a trained supervised classifier for the full study.

➢ *Score Fusion and Risk Classification*

The composite risk score C is computed as a weighted sum of the scaled PHQ-9 score (P) and the NLP linguistic risk score (N):

$$C = 0.65 \times P + 0.35 \times N$$

The higher weighting given to the PHQ-9 stream reflects its status as a validated clinical instrument, while the NLP stream contributes a complementary, passively observable signal. The composite score is mapped to five severity bands.

Minimal (0-19),

Mild (20-39),

Moderate (40-59),

Moderately Severe (60-79), and

Severe (80-100)

Mirroring the banding convention used for PHQ-9 alone, adapted to the 0-100 composite scale.

Table 1 Summarizes the Full Feature Set used by the Fusion Layer.

Feature	Source	Type	Description
PHQ9_Total	Questionnaire	Numeric (0-27)	Sum of the nine PHQ-9 item scores
Sadness_Lexicon_Count	NLP	Count	Frequency of dysphoria-related terms in free text

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

Anhedonia_Lexicon_Count	NLP	Count	Frequency of loss-of-interest phrases
Worthlessness_Lexicon_Count	NLP	Count	Frequency of self-devaluation terms
Self_Harm_Flag	NLP	Boolean	Presence of self-harm / suicidal-ideation language
First_Person_Density	NLP	Ratio	Proportion of first-person pronouns (linked to rumination in prior literature)
Composite_Score	Fusion layer	0-100	0.65 x PHQ-9(scaled) + 0.35 x NLP score

Table 1 Feature set used by the classification / fusion layer.

➢ *Database Design*

To support longitudinal tracking and cohort analytics, the system is architected around a relational schema comprising five tables: participants, phq9_responses, text_entries, nlp_features, and classifications. The schema separates raw clinical responses, raw text, derived linguistic features, and final classification outputs, allowing the feature-extraction and classification logic to be revised independently of the stored raw data. The core schema is shown below.

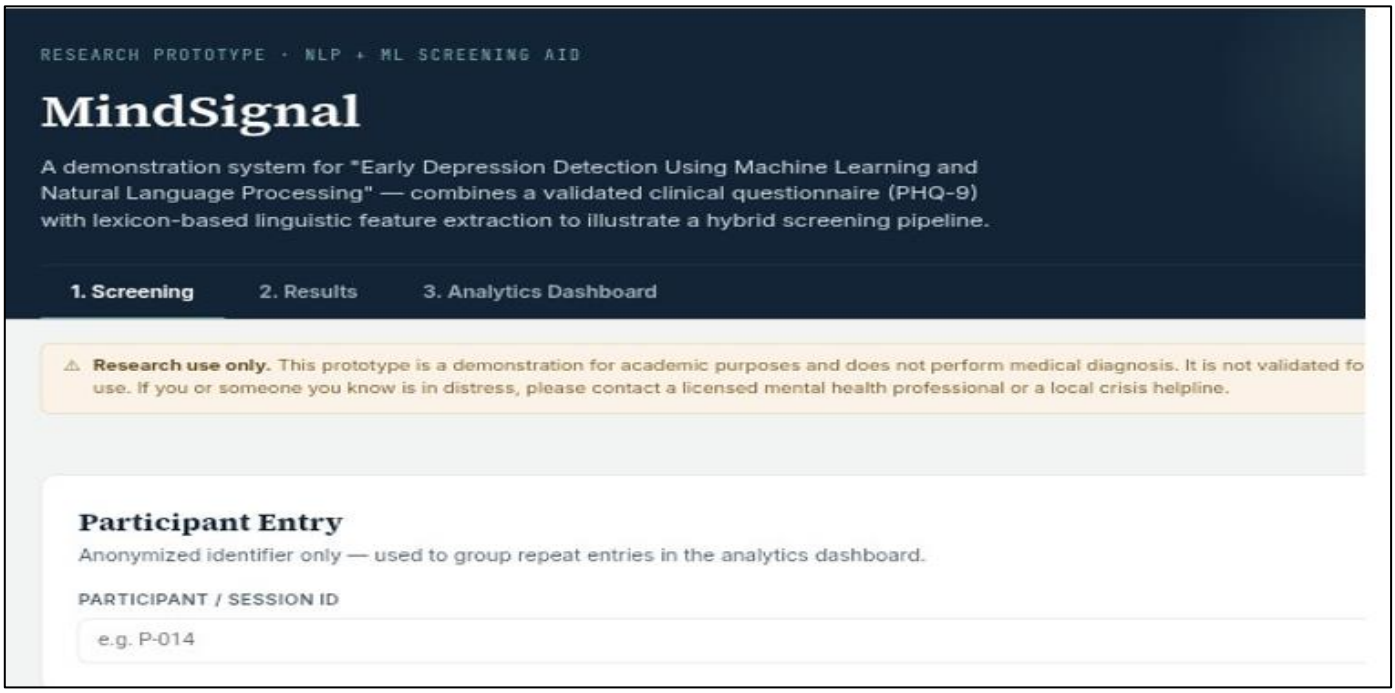
```
CREATE TABLE participants (
  participant_id  VARCHAR(20) PRIMARY KEY,
  created_at      DATETIME DEFAULT CURRENT_TIMESTAMP
);

CREATE TABLE phq9_responses (
  response_id     INT AUTO_INCREMENT PRIMARY KEY,
  participant_id  VARCHAR(20) REFERENCES participants(participant_id),
  item_1..item_9  TINYINT,
  total_score     TINYINT,
  submitted_at    DATETIME DEFAULT CURRENT_TIMESTAMP
);

CREATE TABLE text_entries (
  entry_id        INT AUTO_INCREMENT PRIMARY KEY,
  participant_id  VARCHAR(20) REFERENCES participants(participant_id),
  raw_text        TEXT,
  word_count      INT,
  submitted_at    DATETIME DEFAULT CURRENT_TIMESTAMP
);

CREATE TABLE nlp_features (
  entry_id              INT REFERENCES text_entries(entry_id),
  sadness_count         INT,  fatigue_count    INT,
  anhedonia_count       INT,  isolation_count  INT,
  worthlessness_count   INT,  self_harm_count  INT,
  nlp_risk_score        DECIMAL(5,2)
);

CREATE TABLE classifications (
  classification_id  INT AUTO_INCREMENT PRIMARY KEY,
  participant_id      VARCHAR(20) REFERENCES participants(participant_id),
  phq9_scaled_score    DECIMAL(5,2),
  nlp_risk_score        DECIMAL(5,2),
  composite_score        DECIMAL(5,2),
  risk_band                VARCHAR(30),
  classified_at             DATETIME DEFAULT CURRENT_TIMESTAMP
);
```

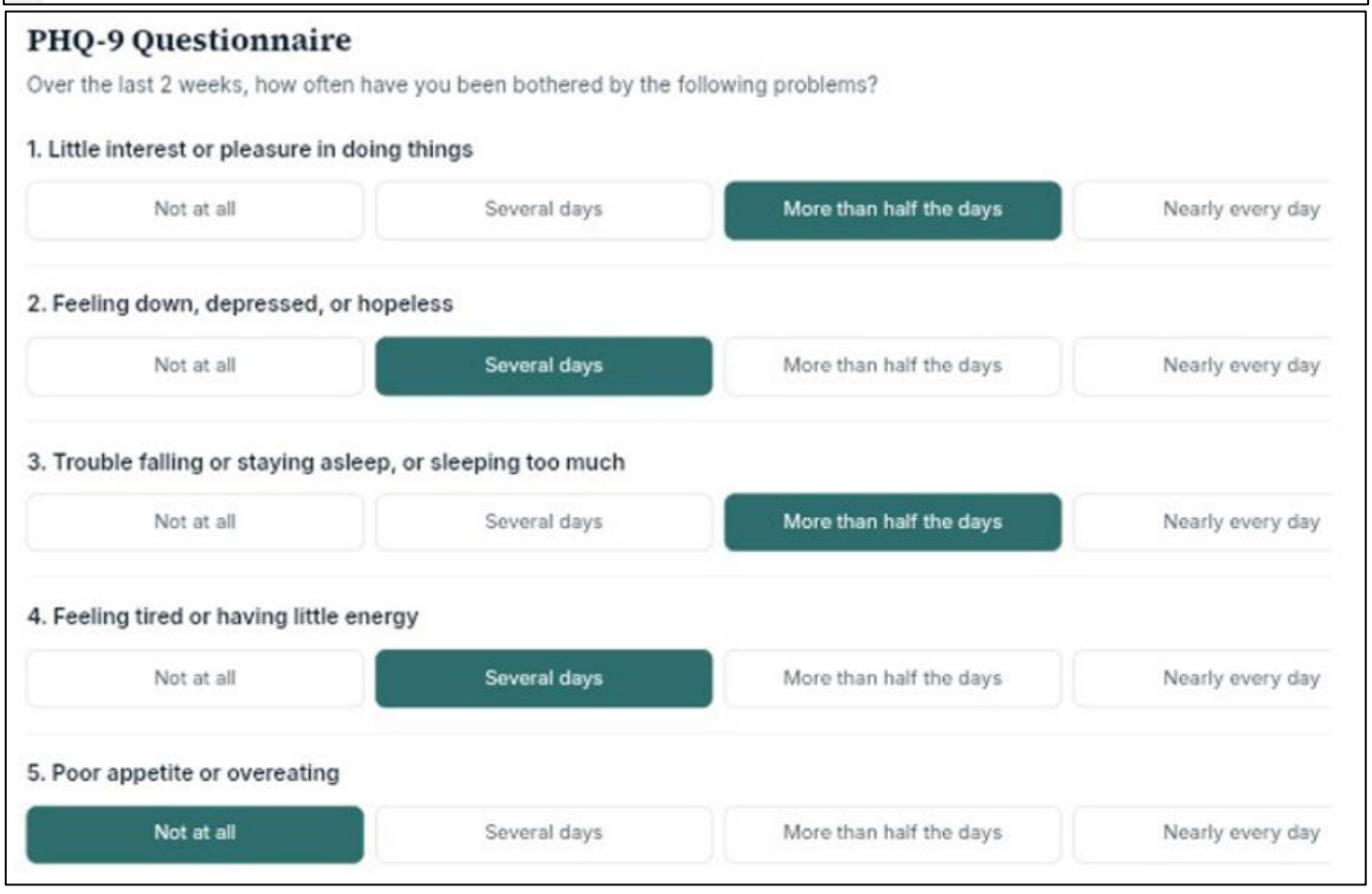
Fig 1 Relational Database Schema Underlying the Screening System.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

➢ *Prototype Implementation*

A working prototype, referred to here as the screening system, was implemented as a browser-based application comprising three user-facing views: (1) a Screening view presenting the PHQ-9 items and a free-text journal box; (2) a Results view displaying the composite score as a radial gauge, the severity band, the PHQ-9 item breakdown, and the detected linguistic indicators with matched terms; and (3) an Analytics Dashboard aggregating all stored screenings into summary statistics, a risk-band distribution chart, and a composite-score trend chart. [Insert here: implementation environment, e.g., "The client was implemented in HTML/JavaScript with Chart.js for visualization, and data was persisted to a relational store (schema in Section 3.5) / to browser-local storage for the prototype phase." Adjust to match your actual deployment.]





ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

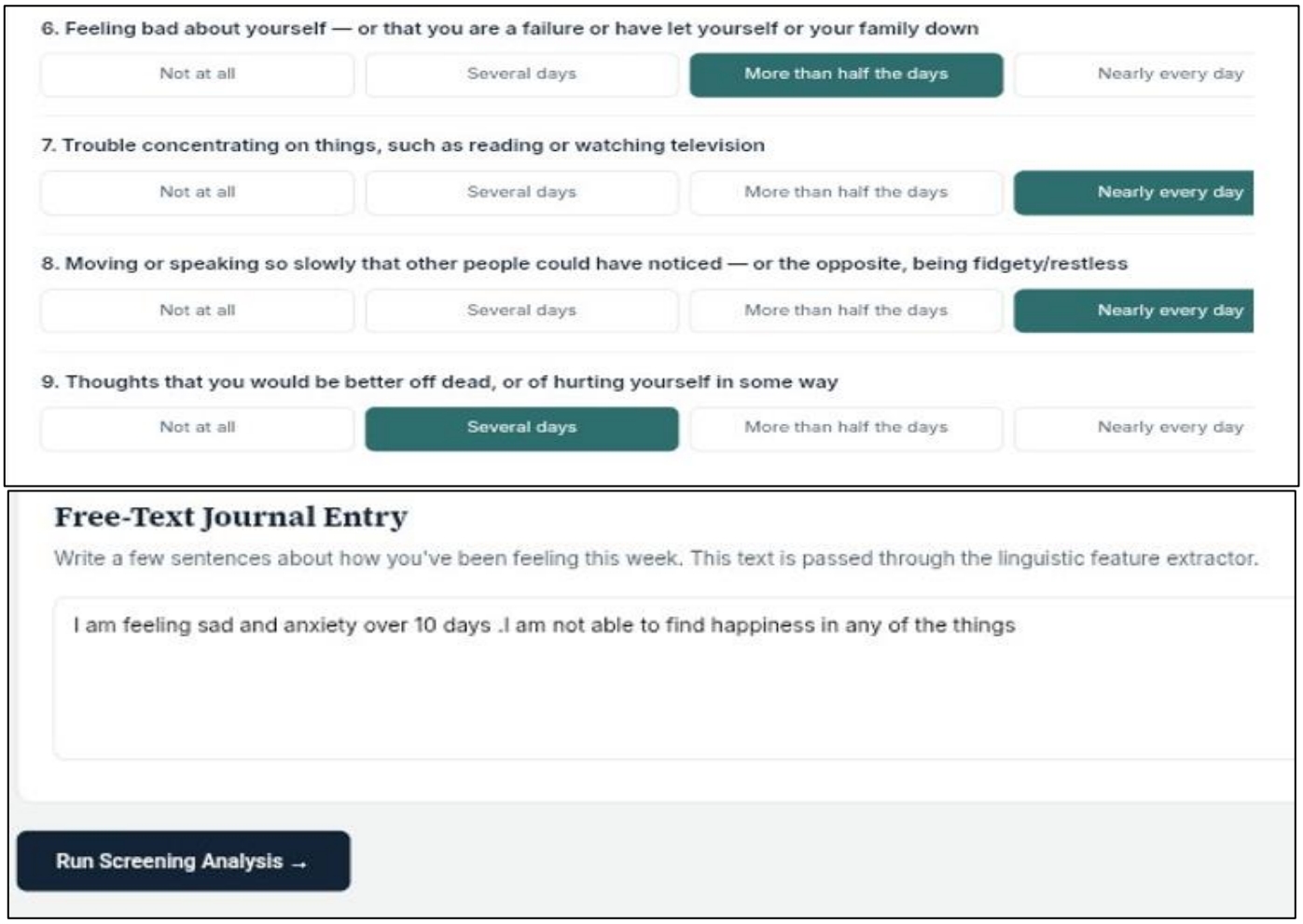
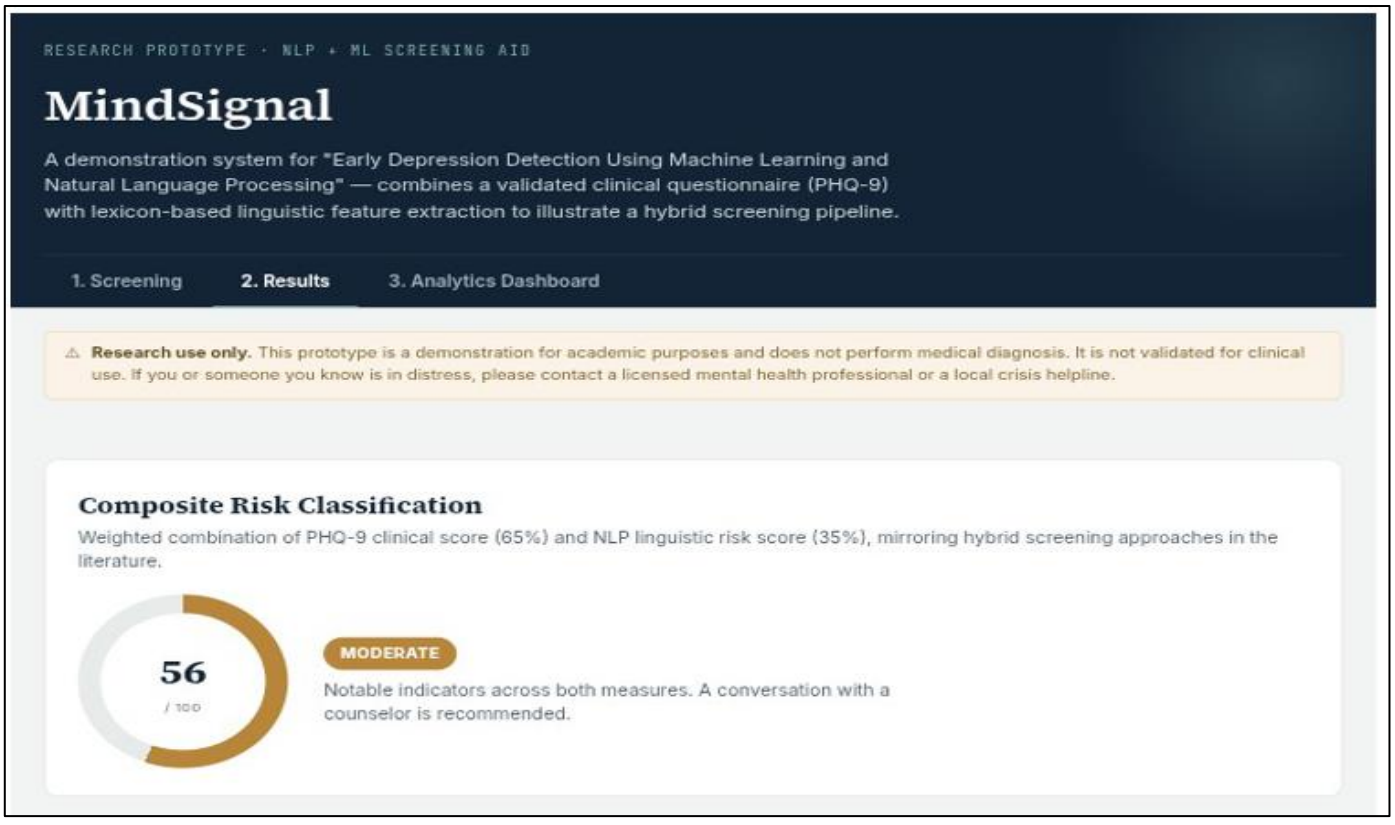


Fig 2 Screening Interface (PHQ-9 and Free-Text Entry).



ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

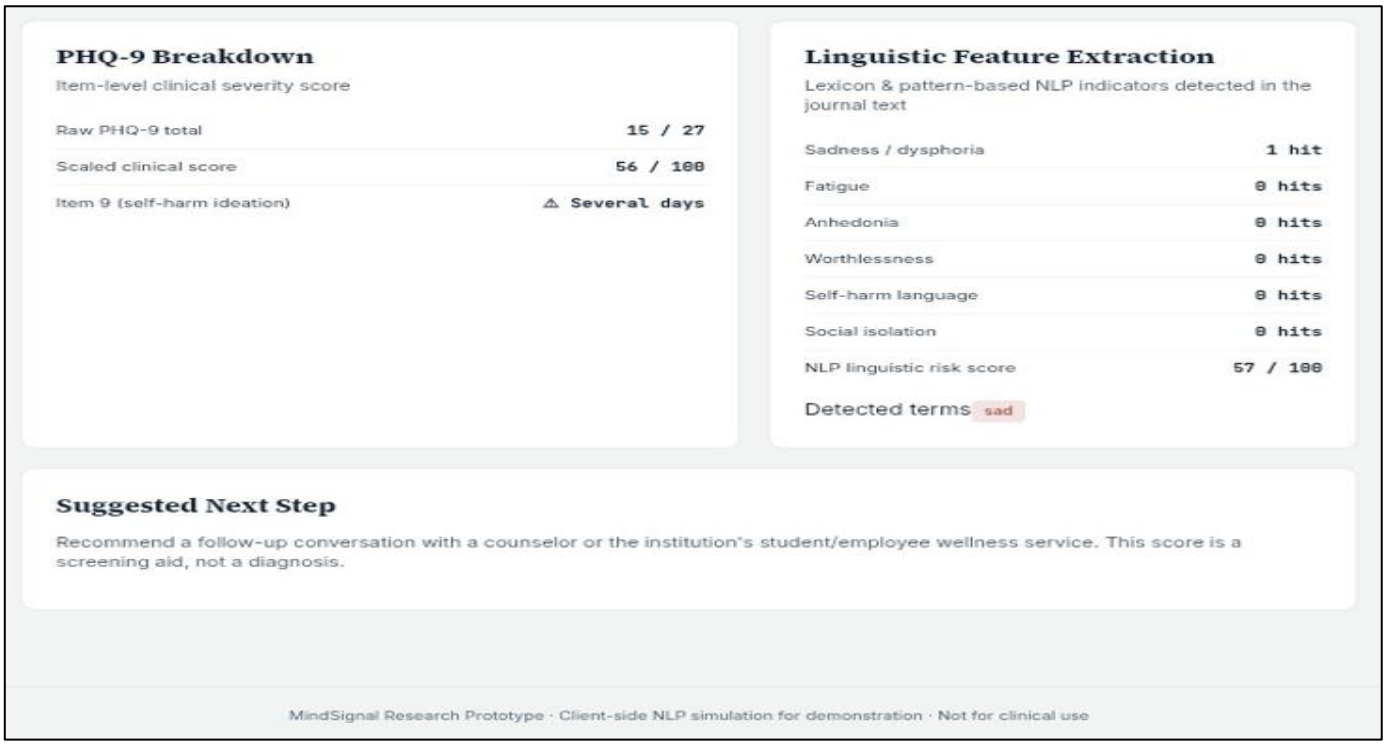
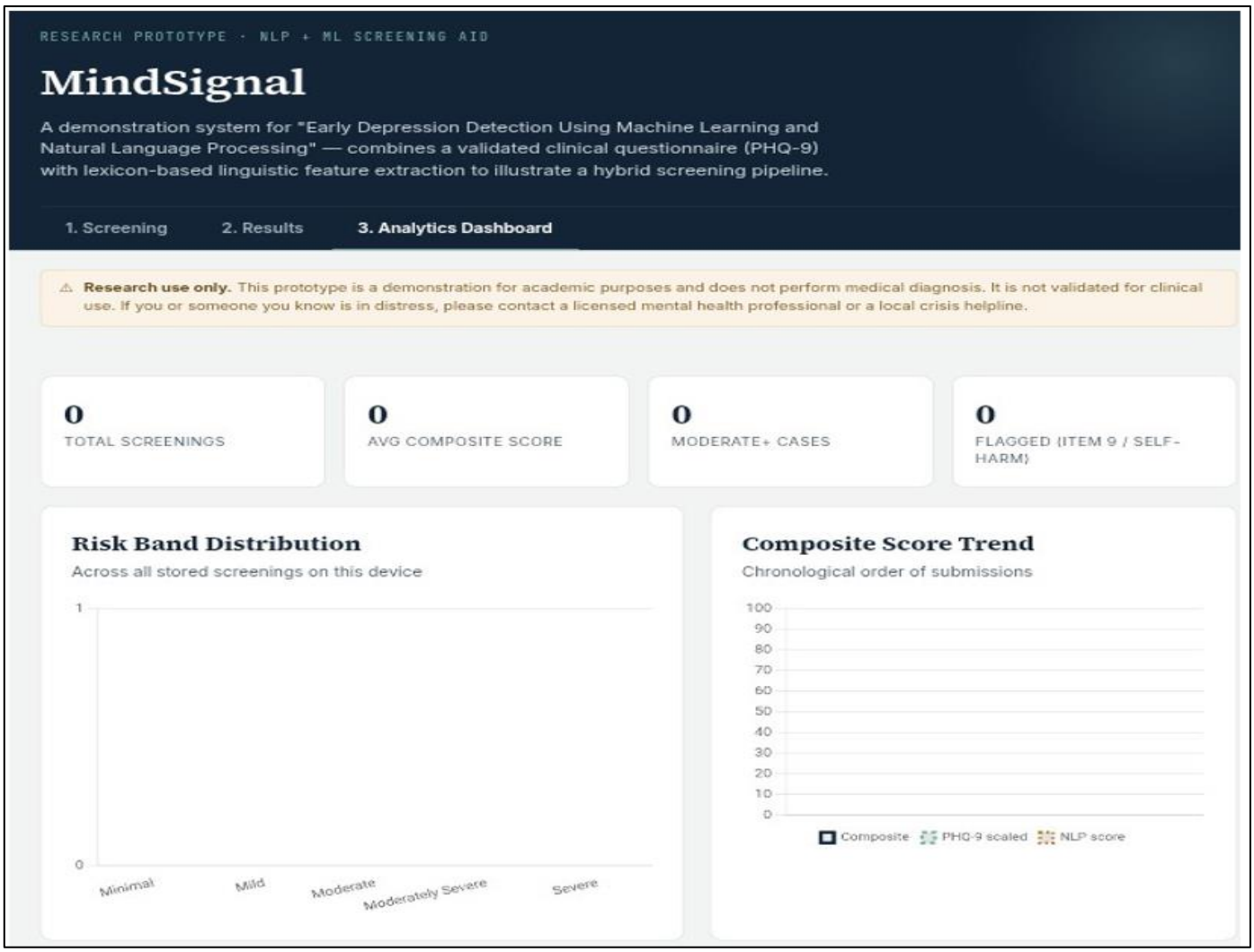


Fig 3 Results View with Composite Risk Gauge and NLP Feature Breakdown.



ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

Fig 4 Analytics Dashboard with Risk-Band Distribution and Score Trend.

IV. RESULTS AND DISCUSSION

[This section should report results from your own pilot data collection. Replace the placeholder text below once you have run the system with a sample of participants or a labeled dataset.]

For an initial pilot with N = [insert number] participants / synthetic test entries, the mean composite score was [insert value] (SD = [insert value]), with [insert number] entries ([insert %]) falling in the Moderate-or-above band. Item 9 (self-harm ideation) was flagged in [insert number] entries. [If you evaluate the NLP module against ground-truth labels, report accuracy, precision, recall, F1-score, and AUC here, and compare against the PHQ-9-only baseline and against the related work summarized in Section 2, e.g., the 99.74% SVM accuracy reported by Motade et al. on PHQ-9-only data [2].]

The dual-stream design allows the two scores (PHQ-9-scaled and NLP-derived) to be compared directly, which is useful diagnostically: cases where the two streams diverge substantially (e.g., a low PHQ-9 score but a high NLP score, or vice versa) may indicate under-reporting on the structured questionnaire or a linguistic pattern not well captured by the lexicon, and could be flagged for closer review in a clinical deployment.

V. LIMITATIONS AND FUTURE WORK

The linguistic feature-extraction module in the current prototype is a transparent, lexicon-based baseline rather than a trained statistical or deep-learning classifier, and its category weights were set heuristically rather than learned from labeled data. The most direct extension of this work is to replace this module with a supervised classifier — for example, TF-IDF features with a Support Vector Machine or Random Forest, or a fine-tuned transformer model such as BERT or DistilBERT — trained and evaluated on a labeled corpus such as the DAIC-WOZ interview dataset, a PHQ-9-labeled survey corpus, or an annotated social-media dataset, with standard classification metrics (accuracy, precision, recall, F1, AUC) reported against a held-out test set.

Additional directions for future work include: (i) validating the composite scoring weights (currently 0.65/0.35) against clinician-rated outcomes; (ii) extending the linguistic feature set with contextual embeddings and sentiment models rather than keyword lexicons alone; (iii) conducting a longitudinal study to assess whether composite-score trends predict clinically meaningful transitions in severity over time; and (iv) an ethical and regulatory review, since any tool of this kind is a screening aid and not a diagnostic device, and would require appropriate clinical oversight, informed consent, and data-privacy safeguards before deployment beyond a research context.

VI. CONCLUSION

This paper presented a hybrid framework for early depression screening that combines the PHQ-9 clinical questionnaire with NLP-based linguistic feature extraction from free-text entries, fused into a single composite risk score with an interpretable breakdown. A working prototype implementing the screening interface, feature-fusion logic, relational database schema, and analytics dashboard was described. While the current linguistic module is a transparent rule-based baseline, the architecture is designed to accommodate a trained ML/NLP classifier in its place, and the paper outlines the evaluation steps required to do so. The framework is intended to support, and not replace, clinical judgment in the early identification of depressive symptoms.

REFERENCES

[1]. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs and Social Media (ICWSM), 2013, pp. 128-137.

[2]. S. Motade, A. Hassan, F. Mir, and K. Parikh, "Machine Learning-Based Approach for Depression Detection Using PHQ-9 and Twitter Dataset," in Proc. 3rd Int. Conf. Communication, Computing and Electronics Systems, Lecture Notes in Electrical Engineering, vol. 844, Springer, Singapore, 2022.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug324

[3]. Soumya Choudhary, Nikita Thomas, Janine Ellenberger, Grish Srinivasan , "A Machine Learning Approach for Detecting Digital Behavioral Patterns of Depression Using Nonintrusive Smartphone Data: Prospective Observational Study," JMIR Formative Research, 2022.

[4]. Hadar Fisher, Nigel M. Jaffer , Kristina Pidcirny ,Anna O. Tierney Mia S. Vaidean , Poorvesh Dongre, and Christian A. Webb, "Language-based detection of depression with machine learning: systematic review and meta-analysis," npj Digital Medicine, 2026.

[5]. Z. Shao, X. Wang, Z. Liu, C. Wang, and K. P. Subbalakshmi, "Systematic Evaluation of Machine-Generated Reasoning and PHQ-9 Labeling for Depression Detection Using Large Language Models," arXiv:2505.17119, 2025.