

Early Detection of Mental Health and Burnout Indicators Using Machine Learning, Transformers, and Explainable AI (XAI)

Keerthika Kumar

Roll No: TDS2627014

Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous)
Sion (W), Mumbai, India

keerthikakumar2512@gmail.com

Project Guide: Asst. Prof. Smita Mangesh Junnarkar

Abstract—Mental health disorders and occupational burnout are major public health concerns that frequently go undetected until they cause severe psychological or physical impairment. Conventional screening relies on self-reported questionnaires and clinical interviews, which are reactive, prone to reporting bias and stigma, and difficult to scale. This paper proposes a hybrid framework for the early detection of mental health and burnout indicators from textual, behavioral, and physiological signals. Transformer-based language models extract semantic and affective cues from text (e.g., journal entries, social media posts, workplace communication); these representations are fused with structured behavioral features such as sleep patterns, activity levels, and communication frequency before classification by classical and ensemble machine learning models. To address the black-box nature of deep models and foster clinical trust, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide global and instance-level interpretations of each prediction. The pipeline is evaluated on public benchmark text data and simulated behavioral data; [insert headline result, e.g., best model with F1 and AUC-ROC]. The framework is intended as a transparent, human-in-the-loop screening aid for clinical and organizational settings.

Keywords—mental health detection, burnout prediction, machine learning, transformers, natural language processing, explainable AI, SHAP, LIME, multimodal fusion

I. INTRODUCTION

Depression, anxiety, and chronic occupational burnout are among the leading causes of disability worldwide [1]. The global burden of mental illness has grown with rising workplace stress, social isolation, and the after-effects of the COVID-19 pandemic. Burnout, recognized by the World Health Organization as an occupational phenomenon, is characterized by emotional exhaustion, depersonalization, and reduced personal accomplishment [2, 3]; left undetected, it can progress to clinical depression and anxiety disorders.

Conventional detection depends on self-report instruments such as the Patient Health Questionnaire (PHQ-9) [4], the Generalized Anxiety Disorder scale (GAD-7) [5], and the Maslach Burnout Inventory (MBI) [2], together with clinical interviews. Although validated, these approaches have three limitations: (i) they are administered periodically rather than continuously and therefore miss early warning signs; (ii) they are susceptible to social desirability bias and under-reporting due to stigma; and (iii) they require active participation, which limits scalability across large populations such as students or employees.

Digital communication, wearable sensors, and organizational productivity tools now make it possible to monitor behavioral and linguistic correlates of well-being passively and continuously. Machine learning (ML) models can learn patterns from such multimodal data, and transformer architectures such as BERT [6] and RoBERTa [7] capture nuanced semantic and affective signals in unstructured text. However, the lack of interpretability of these models is a critical barrier in sensitive domains where predictions may influence clinical or managerial decisions.

This paper proposes an integrated framework combin-

ing (1) transformer-based feature extraction from text, (2) classical and ensemble classifiers trained on structured behavioral and physiological features, and (3) post-hoc explainability through SHAP [8] and LIME [9]. The main contributions are:

1. A hybrid architecture that fuses transformer-derived text embeddings with structured behavioral features for multimodal prediction of mental health and burnout risk.
2. A comparative evaluation of multiple ML and deep learning classifiers on the fused feature space.
3. Integration of global and instance-level explanations to identify the most influential risk indicators.
4. A discussion of ethical, privacy, and deployment considerations for real-world adoption.

The remainder of the paper is organized as follows. Section II reviews related work, Section III describes the methodology, Section IV presents the experimental setup and results, Section V discusses findings, limitations, and ethics, and Section VI concludes with future work.

II. RELATED WORK

A. Machine Learning for Mental Health Detection

Early computational approaches applied Support Vector Machines (SVM) [10], Random Forests [11], and Logistic Regression to structured survey data and engineered linguistic features such as word counts, LIWC categories, and sentiment scores. Studies on social media text showed that depression and anxiety signals could be detected with moderate success [12], but reliance on handcrafted features limited the capture of contextual and semantic nuance.

B. Burnout Prediction

Burnout prediction research has mainly used structured organizational and behavioral data, including working hours, email response patterns, calendar density, wearable sleep and activity data, and periodic self-assessments such as the MBI. Ensemble methods, notably Random Forests and gradient boosting [13], are widely used because of their robustness on tabular data and their ability to handle heterogeneous feature types.

C. Transformer-Based Approaches

BERT [6] and its variants (RoBERTa [7], DistilBERT, and domain-adapted models such as MentalBERT [14]) have substantially improved the detection of affective and cognitive distress markers in text. Through self-attention [15], these models capture long-range context and identify subtle linguistic markers of depression, anxiety, and burnout more accurately than bag-of-words or static embeddings such as Word2Vec and GloVe.

D. Explainable AI in Healthcare

As models grow more complex, interpretability becomes essential in healthcare-adjacent domains where predictions carry ethical and clinical weight. SHAP [8], grounded in cooperative game theory, provides consistent global and local feature attributions, whereas LIME [9] approximates model behavior locally with an interpretable surrogate. Both are increasingly used in clinical decision support to reconcile the trade-off between predictive performance and transparency.

E. Research Gap

Prior work has explored ML-based burnout prediction, transformer-based text analysis, and XAI in clinical contexts largely in isolation. Few studies integrate all three into a unified, multimodal, and interpretable pipeline targeting the joint early detection of mental health decline and occupational burnout. This paper addresses that gap.

III. PROPOSED METHODOLOGY

A. System Overview

The framework (Fig. 1) has four stages: (1) multimodal data acquisition and preprocessing; (2) feature extraction using transformer embeddings and structured behavioral features; (3) feature fusion and classification; and (4) explanation generation with SHAP and LIME.

B. Data Sources

The framework operates on three data types:

- *Textual data*: journal entries, anonymized social media posts, chat or communication logs, and open-ended survey responses.
- *Structured behavioral data*: sleep duration and quality, physical activity, screen time, communication frequency, and calendar or meeting density.
- *Self-report labels*: ground-truth or proxy labels from validated instruments, namely PHQ-9 (depression), GAD-7 (anxiety), and MBI (burnout), used for supervised training.

Where real clinical data are unavailable owing to privacy constraints, public benchmark datasets (e.g., DAIC-WOZ [16], Dreddit [17], and CLPsych shared-task datasets) and

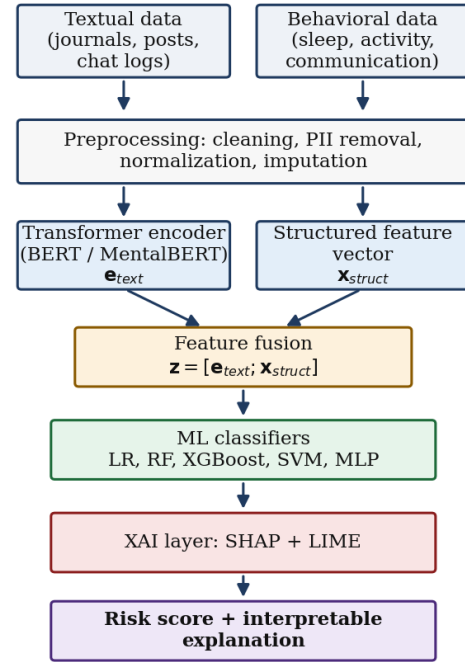


Figure 1: Architecture of the proposed multimodal early-detection pipeline.

simulated behavioral data are used for experimentation and validation.

C. Preprocessing

Text is tokenized, lowercased, stripped of personally identifiable information (PII), and filtered for noise such as URLs and special characters. Structured behavioral features are scaled with min-max normalization, and missing values are imputed by the median or by model-based imputation (e.g., k -nearest neighbors), depending on the missingness pattern.

D. Transformer-Based Text Feature Extraction

A pre-trained transformer (BERT-base or a domain-adapted variant such as MentalBERT) encodes each text. The final hidden state of the [CLS] token, or a mean-pooled representation of the token embeddings, is extracted as a dense vector $\mathbf{e}_{\text{text}} \in \mathbb{R}^d$, where $d = 768$ for BERT-base.

E. Feature Fusion

The text embedding is concatenated with the normalized structured feature vector $\mathbf{x}_{\text{struct}} \in \mathbb{R}^m$ to give the unified representation

$$\mathbf{z} = [\mathbf{e}_{\text{text}}; \mathbf{x}_{\text{struct}}] \in \mathbb{R}^{d+m}, \quad (1)$$

which is the input to the downstream classifiers and lets the system reason jointly over linguistic and behavioral evidence.

F. Classification Models

The following classifiers are trained and compared on $\mathbf{z}$:

- *Logistic Regression (LR)*: an interpretable baseline.
- *Random Forest (RF) and gradient boosting (XGBoost)*: robust on heterogeneous, tabular-style features.
- *SVM with RBF kernel*: non-linear decision boundaries.
- *Multi-layer perceptron (MLP)*: a shallow feed-forward

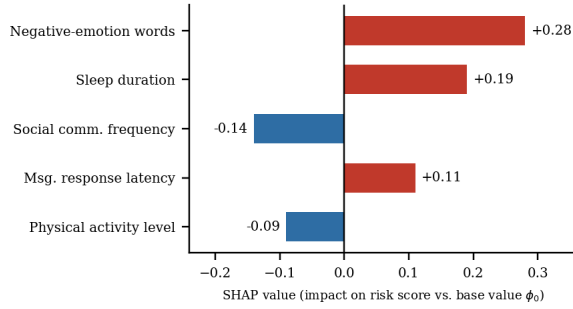


Figure 2: Illustrative (not experimental) SHAP attribution for one prediction. Red bars raise and blue bars lower the risk score relative to the base value ϕ_0 .

network on $\mathbf{z}$.

Each model outputs a probability $\hat{y} \in [0, 1]$ estimating the risk of mental health decline or burnout, which can be thresholded for binary classification or retained as a continuous score for triage.

G. Explainable AI Layer

1) *SHAP*. SHAP attributes a model’s output for each prediction to individual input features using Shapley values from cooperative game theory. For a model f and input $\mathbf{z}$, the attributions ϕ_i satisfy the additivity property

$$f(\mathbf{z}) = \phi_0 + \sum_{i=1}^n \phi_i, \quad (2)$$

where ϕ_0 is the base value (mean model output) and $n = d + m$ is the number of fused features. Summary and dependence plots identify globally important indicators such as specific linguistic markers, sleep disruption, or reduced communication.

2) *LIME*. For instance-level explanations, LIME fits a sparse interpretable surrogate (e.g., linear regression) on perturbed samples around a given input, yielding statements such as “elevated risk driven primarily by reduced sleep duration and increased use of negatively valenced language.” Fig. 2 illustrates the form of a SHAP attribution for a single prediction.

H. Deployment Considerations

The pipeline is designed for a dashboard used by clinicians, counselors, or HR wellness teams that presents the risk score with its SHAP/LIME explanation, supporting human-in-the-loop decisions rather than automated diagnosis.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Datasets

Experiments combine a public text-based mental health dataset with simulated structured behavioral data to emulate a multimodal setting. [Insert dataset name, number of samples, class distribution, train/test protocol, and how the simulated behavioral features were generated.]

B. Implementation and Evaluation Metrics

[Insert transformer checkpoint, maximum sequence length, pooling, hyperparameters, cross-validation scheme, random seed, and hardware.] Performance is reported using accuracy, precision, recall, F1-score, and area under the

Table 1: Comparative performance of classification models

Model	Acc.	Prec.	Rec.	F1	AUC
Logistic Regression	—	—	—	—	—
SVM (RBF)	—	—	—	—	—
Random Forest	—	—	—	—	—
XGBoost	—	—	—	—	—
MLP (proposed)	—	—	—	—	—

Table 2: Ablation of input modalities (Logistic Regression)

Input	Acc.	Prec.	Rec.	F1	AUC
Text only	—	—	—	—	—
Behavior only	—	—	—	—	—
Fused	—	—	—	—	—

ROC curve (AUC-ROC). Recall receives particular attention because of the clinical cost of false negatives, i.e., missed at-risk individuals.

C. Baseline Comparison

Table 1 compares the evaluated models (mean $\pm$ standard deviation over cross-validation folds).

D. Ablation Study

To quantify the contribution of each modality, a text-only, a behavior-only, and the fused model are compared (Table 2).

E. Explainability Analysis

SHAP global importance rankings are examined for the linguistic markers (e.g., first-person singular pronoun frequency, negative-emotion word density) and behavioral markers (e.g., reduced sleep duration, decreased social communication) that contribute most to elevated risk, and LIME explanations are inspected for representative correctly and incorrectly classified cases. [Insert SHAP summary plot, LIME instance explanation, and a discussion of the observed patterns.]

V. DISCUSSION

A. Key Findings

Linguistic and behavioral signals capture complementary aspects of psychological state, so fusing transformer representations with structured behavioral indicators is hypothesized to outperform unimodal baselines. [Confirm or revise this statement using Tables 1 and 2.] Explainability analysis adds actionable insight beyond a raw risk score, which may increase clinician and end-user trust and enable earlier intervention.

B. Limitations

- *Data scarcity and labeling*: high-quality labeled multimodal mental health datasets are scarce because of privacy and ethical constraints, limiting generalizability.
- *Simulated behavioral data*: pairing real text with simulated behavioral features does not reflect true within-person relationships, so multimodal gains must be validated on real paired data.
- *Bias and fairness*: language and behavior vary across demographic and cultural groups; models trained on

non-representative data risk biased predictions.

- *Correlation versus causation*: detected indicators are correlational, not diagnostic; the system is a screening aid, not a clinical diagnostic tool.

C. Ethical and Privacy Considerations

Given the sensitivity of mental health data, deployment must comply with data protection standards (e.g., HIPAA, GDPR, and applicable Indian data protection law), use informed consent, anonymize textual data, and avoid punitive use of burnout predictions in organizations. Human oversight is essential, and the system should be positioned strictly as decision support and early screening rather than an autonomous diagnostic authority.

VI. CONCLUSION AND FUTURE WORK

This paper presented a hybrid ML and transformer-based framework, augmented with SHAP and LIME, for the early detection of mental health and burnout indicators from textual and behavioral data. By pairing the representational power of transformers with transparent explanations, the approach aims to balance predictive performance with the interpretability required for responsible clinical and organizational use. Future work will (i) validate the framework on larger, longitudinal, real-world multimodal datasets; (ii) incorporate wearable physiological signals such as heart rate variability; (iii) extend the framework to continual and federated learning to preserve privacy; and (iv) conduct user studies with clinicians and HR professionals to assess the utility and trustworthiness of the explanations.

ACKNOWLEDGMENT

The author thanks the Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous), Sion (W), Mumbai, for the academic environment and support for this work, and the project guide, Asst. Prof. Smita Mangesh Junnarkar, for her guidance and encouragement.

REFERENCES

- [1] World Health Organization, "Mental disorders," WHO Fact Sheets, 2022.
- [2] C. Maslach and S. E. Jackson, "The measurement of experienced burnout," *J. Organizational Behavior*, vol. 2, no. 2, pp. 99–113, 1981.
- [3] C. Maslach, W. B. Schaufeli, and M. P. Leiter, "Job burnout," *Annu. Rev. Psychol.*, vol. 52, pp. 397–422, 2001.
- [4] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.
- [5] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, 2006.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [7] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
- [8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.
- [9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
- [10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop on Computational Linguistics and Clinical Psychology*, 2014, pp. 51–60.
- [13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.
- [14] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proc. LREC*, 2022.
- [15] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
- [16] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proc. LREC*, 2014, pp. 3123–3128.
- [17] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *Proc. LOUHI Workshop*, 2019, pp. 97–107.