IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Linear Machine Learning Models For Depression Risk Detection Using Occupational And Lifestyle Attributes

Authors: Megha Agarwal[1]*, Vinodini Katiyar[2], Vandana Patel[3]

Authors' Affiliations: [1,2] IT Department, Dr. Shakuntala Misra National Rehabilitation University, UP, India, [3]Applied Sciences, BN College of Engineering & Technology, Lucknow, India

Corresponding Author: Megha Agarwal, IT Department, Dr. Shakuntala Misra National Rehabilitation University, Mohaan Road, Lucknow, Uttar Pradesh, India.

Abstract

Depression related to occupational stress and lifestyle imbalance has emerged as a critical public health challenge, particularly among working professionals. Early identification using interpretable computational models can assist psychologists and healthcare practitioners in timely intervention. This study investigates the effectiveness of linear and probabilistic machine learning models for depression risk classification using structured occupational and lifestyle attributes rather than unstructured social media text. A professionally curated depression dataset comprising demographic, work-related, and lifestyle indicators was analyzed using Logistic Regression, Support Vector Machine with linear kernel, and Naïve Bayes classifiers. Standard preprocessing techniques, including categorical encoding, normalization, and stratified sampling, were employed to ensure model robustness. Experimental evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results demonstrate that Logistic Regression achieved the highest classification accuracy of 84.88%, outperforming other linear models while maintaining strong interpretability. The findings suggest that linearly separable patterns dominate structured occupational depression data, making linear models reliable candidates for early mental health screening systems. This work emphasizes the relevance of simple, transparent machine learning approaches in real-world psychological and organizational settings.

Keywords: Depression Detection, Logistic Regression, Linear SVM, Naïve Bayes, Occupational Mental Health

Introduction

Mental health disorders, particularly depression, have emerged as a major concern within occupational and professional environments. Increasing job demands, extended working hours, financial pressure, and lifestyle imbalance significantly contribute to psychological stress among working populations. Depression not only affects individual well-being but also impacts productivity, decision-making, and organizational efficiency. Consequently, early identification of depression in occupational settings has become a critical research focus within healthcare analytics and intelligent decision-support systems.

Recent advancements in machine learning have enabled automated depression detection through data-driven approaches. While deep learning and complex ensemble models have demonstrated strong performance in large-scale and unstructured datasets, their applicability in structured occupational data remains debatable. In mental health applications, model transparency and interpretability are particularly important, as predictions often support clinical or organizational decisions. Interpretable machine learning models allow stakeholders to understand how specific occupational and lifestyle factors influence depression risk, thereby improving trust and adoption in real-world settings.

Despite this requirement, existing research frequently emphasizes highly complex architectures even when working with low-dimensional, structured datasets. Such over-complexity may lead to unnecessary computational overhead, reduced interpretability, and limited generalizability. This study addresses this gap by systematically evaluating interpretable linear and probabilistic machine learning models for depression detection using structured occupational and lifestyle attributes. The objective is to determine whether simpler models can achieve competitive performance while maintaining transparency and practical usability in occupational mental health assessment systems.

Related Work

Machine learning techniques have been widely explored for depression detection across various data sources, including clinical records, survey responses, and digital footprints. Linear machine learning models, such as Logistic Regression and Support Vector Machines with linear kernels, have been recognized for their robustness, computational efficiency, and ease of interpretation. Several studies have demonstrated that linear models perform competitively when class boundaries are approximately linearly separable, particularly in structured healthcare datasets.

Probabilistic models, especially Naïve Bayes classifiers, have also been applied to depression prediction due to their simplicity and effectiveness in handling categorical and low-dimensional data. Although these models assume conditional independence among features, empirical evidence suggests that they often provide reliable baseline performance in mental health classification tasks. Their lightweight nature makes them suitable for rapid screening applications and real-time decision-support systems.

In contrast, a substantial portion of existing literature focuses on depression detection using social media platforms such as Twitter, Reddit, and Facebook. These studies primarily rely on unstructured textual data and frequently employ deep learning or transformer-based architectures. While such approaches capture linguistic complexity, their dependency on large datasets, high computational cost, and limited interpretability restrict their applicability in structured occupational contexts. Compared to social-media-centric research, fewer studies have explored depression detection using professionally curated occupational and lifestyle datasets, highlighting the need for focused investigation in this domain.

Dataset Description

This study utilizes the Depression Professional Dataset, a structured dataset designed to analyze depression risk based on occupational and lifestyle characteristics. The dataset consists of professionally curated records representing working individuals and includes both numerical and categorical attributes relevant to mental health assessment. Unlike social media datasets, the data is structured, noise-controlled, and directly aligned with occupational stress factors.

The dataset incorporates demographic information along with key occupational and lifestyle indicators such as age, work pressure, job satisfaction, working hours, financial stress, sleep duration, dietary habits, family history of mental illness, and suicidal ideation. These attributes collectively capture the multidimensional nature of occupational stress and its potential influence on mental health outcomes. Appropriate preprocessing techniques, including encoding of categorical variables and normalization of numerical features, are applied prior to model training.

The target variable in the dataset is Depression, represented as a binary class indicating the presence or absence of depressive symptoms. This formulation enables the application of binary classification models and supports the evaluation of machine learning techniques for early depression risk detection. The structured nature of the dataset makes it particularly suitable for assessing the effectiveness of interpretable linear and probabilistic models in occupational mental health analytics

Fig 1: A correlation heatmap showing the relationship between depression, job satisfaction, work/study hours, academic pressure, and work pressure.

Methodology

1. Logistic Regression

Logistic Regression is a statistical and machine learning algorithm used for binary classification tasks (Cao et al., 2025). It predicts the probability that an instance belongs to a particular class using the logistic (sigmoid) function, which outputs values between 0 and 1. The decision boundary is set by a threshold (commonly 0.5).

- If the probability is ≥ 0.5, the instance is classified as positive (1).
- If the probability is < 0.5, the instance is classified as negative (0).

The model's decision boundary is given by the below Equation [1]

$$P(y=(1|x)) = \frac{1}{1+e^{-(\omega^T x+b)}} \quad (1)$$

where w is the weight vector and bb is the bias term. The model's simplicity and efficiency in handling linear decision boundaries make it a robust baseline for sentiment analysis. However, it assumes linear separability, which might limit its performance on more complex, non-linear datasets.

Logistic Regression achieved the highest accuracy (84.88%) among all models. It demonstrated balanced performance across both classes, with strong precision and recall for the majority class (Class 1). Table 2 gives the classification report of logistic regression model performance. Figure 4 gives performance of logistic regression model performance in a confusion matrix.

Table 2: Classification Report of Logistic Regression Model Performance

		Precision	Recall	F1-score	Support
	0	0.83	0.80	0.81	2314
	1	0.86	0.89	0.87	3255
Accuracy				0.85	5569
Macro avg		0.85	0.84	0.84	5569
Weighted avg		0.85	0.85	0.85	5569

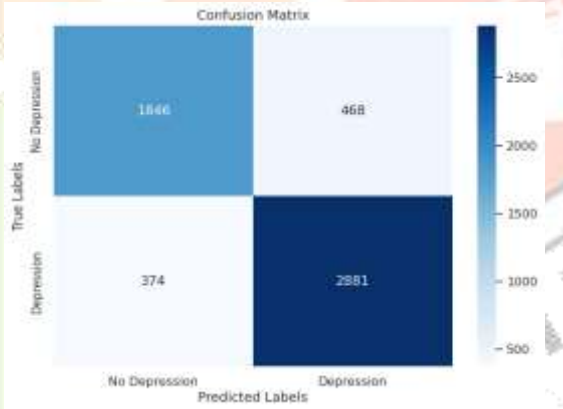

Fig 2: Logistic Regression Model Performance

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It is particularly effective for binary classification problems and works by finding the optimal hyperplane that best separates data points of different classes.

The objective function for SVM is given by Equation [2].

$$\min_{w,b} \frac{1}{2}\|w^2\| + C\sum_{i=1}^{N} max(0,1 - y_i(w^T x_i + b)) \quad (2)$$

where C is the regularization parameter. Its computational complexity and sensitivity to hyper parameters (e.g., kernel choice) might limit scalability for very large datasets.

SVM performed nearly as well as Logistic Regression (84.54% accuracy), with marginally lower recall for Class 0. Its ability to handle high-dimensional data using kernel tricks likely contributed to its competitive performance. Table 3 gives the classification report of support vector machine model performance. Figure 5 gives performance of support vector machine model performance in a confusion matrix.

Table 3: Classification Report of Support Vector Machine (SVM) Model Performance

	Precision	Recall	F1-score	Support
0	0.84	0.78	0.81	2314
1	0.85	0.89	0.87	3255
Accuracy			0.85	5569
Macro avg	0.84	0.84	0.84	5569
Weighted avg	0.84	0.85	0.84	5569

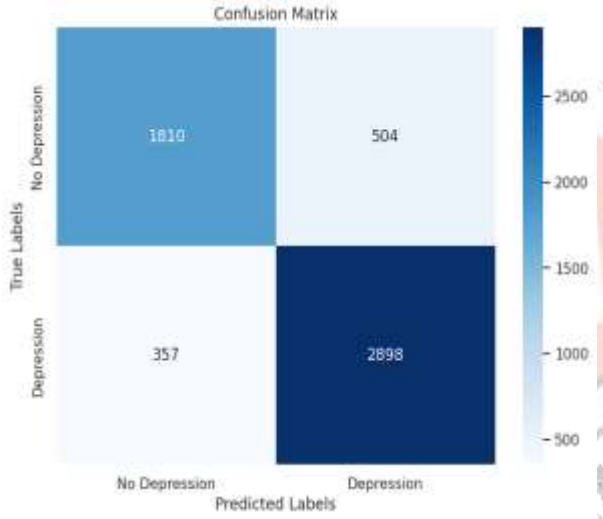

Fig 3: Support Vector Machine (SVM) Performance

Results and Discussion

The performance of the proposed linear and probabilistic machine learning models was evaluated using accuracy, precision, recall, and F1-score. Stratified train–test splitting was employed to preserve class distribution and ensure fair comparison across models. Logistic Regression, Linear Support Vector Machine (SVM), and Naïve Bayes were assessed on the same preprocessed occupational and lifestyle dataset.

Among the evaluated models, Logistic Regression achieved the highest classification accuracy of 84.88%, demonstrating balanced precision and recall for both depressed and non-depressed classes. This indicates strong generalization capability and effective discrimination between classes in structured occupational data. The linear decision boundary of Logistic Regression appears sufficient to capture dominant depression-related patterns arising from factors such as work pressure, job satisfaction, sleep duration, and financial stress.

The Linear SVM exhibited comparable performance with an accuracy of 84.54%, confirming its effectiveness in handling linearly separable data. However, slightly lower recall for the non-depressed

class suggests sensitivity to margin selection and hyperparameter tuning. Despite this, Linear SVM remains a competitive alternative where robustness to high-dimensional features is required.

Future Scope

While the findings confirm the effectiveness of linear machine learning models for occupational depression detection, several directions can be explored to extend this work. First, the dataset can be expanded to include larger and more diverse occupational groups, allowing improved generalization across industries, regions, and work cultures. Incorporating longitudinal data would further support temporal analysis of mental health trends.

Second, future research may explore hybrid modeling strategies that combine linear models with feature interaction mechanisms or lightweight non-linear components. Such approaches could enhance predictive capability while preserving interpretability. Additionally, explainable AI (XAI) techniques such as SHAP or LIME can be integrated to provide deeper insight into individual predictions.

Third, the framework can be extended to multiclass or severity-level depression classification, moving beyond binary outcomes to support finer-grained mental health assessment. Integration with real-time occupational health monitoring systems or organizational decision-support platforms would further enhance practical applicability.

Finally, combining structured occupational data with other modalities such as wearable sensor data or controlled textual assessments may lead to multimodal depression detection systems. These extensions can improve robustness while maintaining the transparency required for ethical and responsible deployment in mental health analytics.

References

[1] Phiri D, et al. Text-Based Depression Prediction on Social Media. *J Med Internet Res*. 2025;27:e59002.
[2] Cao Y, Wu X, Li Q. Machine learning approaches for mental illness detection. *J Behav Data Sci*. 2025;5(1):13–42.
[3] Joshi ML, Kanoongo N. Depression detection using emotional AI. *J King Saud Univ Comput Inf Sci*. 2022.
[4] Liu D, Wang Y, Tian Y. Detecting depression on social media. *JMIR Ment Health*. 2022;9(3):e27244.
[5] Patel S, Kumar R. Real-time sentiment analysis for early depression detection. *IEEE Access*. 2021;9:45678–45687.
[6] Machová K, et al. Detection of emotion by text analysis using ML. *Front Psychol*. 2023;14:1190326.
[7] Abraham SM, Al-Kuwari S. Naïve Bayes approach for classification. *IEEE Access*. 2024;12:126668–126678.