Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 6, pp. 1821–1829
DOI : 10.47974/JIOS-2011

An explainable artificial intelligence-based approach for prediction and analysis of mental health disorder due to work pressure in the tech industry

Vaishali Mehta Wadhwa *
Akanksha Mahajan †
Department of Computer Science and Engineering
Panipat Institute of Engineering and Technology
Panipat 132101
Haryana
India

Tamanna Jena §
Silberman School of Business
Fairleigh Dickinson University
Vancouver V6B2P6
Canada

Jyoti Garg ‡
Department of Computer Science and Engineering
Maharishi Markandeswar Deemed to be University
Ambala 133203
Haryana
India

Abstract

These days, mental health disorders are becoming a major problem, particularly for workers. Effective management and treatment of mental health illnesses depend on

* *ORCID-ID:* `0000-0003-2017-0528`
† *ORCID-ID:* `0000-0003-2775-7751`
§ *ORCID-ID:* `0000-0003-2703-7420`
‡ *ORCID-ID:* `0009-0000-2762-3599`

* *E-mail:* `drvaishaliwadhwa@gmail.com` (Corresponding Author)
† *E-mail:* `er.aakanksha@yahoo.co.in`
§ *E-mail:* `tamannasinghdeo@gmail.com`
‡ *E-mail:* `jyotigarg1055@gmail.com`

early and precise diagnosis. This study investigates how to enhance the interpretability of machine learning models in the prediction of mental illness by utilizing explainable artificial intelligence (XAI) techniques, namely SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). We used a dataset from a survey of IT companies that had the required variables in order to create a prediction model. Our results showed that important variables like age, privacy, and work-related distractions had a big impact on model predictions. SHAP values shed emphasis on the role of individual characteristics, emphasizing the significance of age, work interference, and privacy as key risk factors for mental illness. The dependent graphs showed a strong relationship between key characteristics and model output, highlighting the necessity of employee diagnosis and preventative care. The findings highlight the potential of XAI techniques in therapeutic settings, opening the door to more individualized diagnosis and better outcomes for patients with mental health conditions.

Subject Classification: *68T37.*

Keywords: *Explainable AI, LIME, SHAP, Machine learning, Mental health disorder, Predictive nodeling.*

1. Introduction

Mental health is a global health concern in many countries. 51.5 million Americans, which is about 20% of the 2021 population, had a mental illness as per the National Institute of Mental Health (NIMH) [1]. 31.2% of these people had anxiety, making it the most prevalent condition. The second most prevalent condition is depression, with an estimated annual prevalence of 17.3 million people [2]. IT industry is one such sector were employees suffer from mental health issues due to excessive work pressure [3]. The deployment of AI in healthcare has changed health outcomes in the recent past [4]. One of the benefits is enhanced diagnostic accuracy and therapeutic efficacy [5]. Traditional AI models often act as black boxes, using hidden, convoluted mechanisms to make predictions. Specifically in industries where the stakes are high, like healthcare, banking, and law, a lack of interpretability may lead to doubts and fear [6]. XAI works to narrow the gap by showing how AI systems perform their functions [7]. The primary goal of XAI is to offer some understanding of AI outputs, from the operation of the system to the reasoning that led to the decisions made by these systems [8][9]. According to [10], machine learning or ML algorithms can precisely predict and analyze mental health disorders such as depression, dementia, and schizophrenia through their ability to identify intricate patterns in big data sets and electronic health records, as well as behavioral data and genetics [11].

2. Literature Review

AI can drastically alter mental health services by facilitating early detection and prompt interventions. Machine learning (ML) algorithms have been the subject of numerous studies to identify outcomes associated with mental health and other medical issues. The study in [12] evaluated the patterns of curative response in patients with depression. Machine learning approaches were used to obtain encouraging output replies. Similar to this, signs of mental health issues can be found in social media posts by applying natural language processing (NLP) techniques [13]. Explainability has become crucial, particularly in the healthcare industry, where consumers and professionals will need to understand AI decisions. Explainable models can increase treatment adherence, acceptance, and confidence in the recommended approach [14], [15]. Subjective experiences are crucial, and understanding the rationale behind AI-generated predictions would undoubtedly significantly improve therapy outcomes for a person's mental health [16]. Many AI techniques have been used to forecast mental health issues. Machine learning algorithms have been used to assess data from clinical examinations and wearable devices, such as social media. Deep learning models are more accurate in identifying patterns, especially in large datasets like electronic health records (EHRs) [17]. Additionally, text data from social media and forums has been analyzed for sentiment analysis and to identify potential mental health indicators using natural language processing (NLP) methods [18].

3. Proposed Methodology

The dataset was collected from a 2014 survey that gauges the prevalence of mental health illnesses in IT workplaces as well as attitudes toward mental health. For several features, the average of the two closest observations was substituted over time for the corresponding observation in cases where no information was available. By utilizing scikit-learn's MinMaxScaler. We reduce the number of features in our dataset prior to model creation. PCA transformation methods are used to achieve this. To improvise the model's prediction capabilities and examine the effects of particular mental stressors, we worked with an industry expert to update the collection of input features. Hence, for ML models to produce better results, some of the categorical features have been removed during model development. The proposed work aims to understand the logic behind the specific prediction made by the machine learning model. A Google Colab Pro cloud-based Python environment was used for the simulations.

3.1 *XAI Methods*

We select SHAP in conjunction with feature permutation approaches and LIME for local interpretations. Coalitional game theory elements were introduced to the machine learning domain to create SHAP, a benchmark XAI technique. LIME is a popular algorithm that is a model-agnostic method for generating local interpretations. Although global explanations can still be produced. All of the previously stated XAI techniques have been used on the collected and modified datasets to determine the global feature rankings of the top-performing models.

4. Results and Discussion

1.1 *Feature Importance*

The visual representation of the Random Forest models with and without feature engineering is shown in Figure 1.

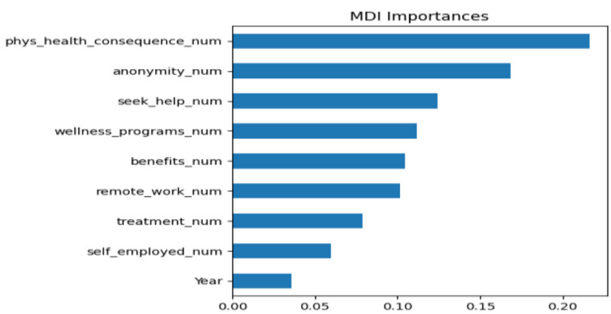

Figure 1

Feature importance

4.2 *LIME (Local Interpretable Model-agnostic Explanations)*

LIME examines the impact of altering the machine learning model's input data on the predictions. Using sample permutations and the matching forecasts from the traditional ML model, LIME produces a new dataset. The interpretable model is then trained using this newly generated dataset. Although the LIME model does not provide a good global approximation, the predictions of the machine learning model must be

Figure 2

LIME interpretations for 8th instance

well-represented locally by the learned model. The LIME explanation for the eighth instance is shown in Figure 2.

The likelihood of no disease is 0.80, and the likelihood of having disease is 0.10 as explained by the LIME model. The predicted classes are positively impacted by the features named mental_vs_physical_yes, mental_vs_physical_no, gender_female, and anonymity_no , while the anticipated classes are negatively impacted by every other variable. By estimating the model's predictions using interpretable surrogate models, LIME gave us local explanations that helped us comprehend the contributions of distinct variables for certain cases.

4.3 *SHAP(Shapley additive explanations)*

In this study, we employed SHAP values to infer the outcomes of our machine learning model for detecting mental health disorders. By clearly illustrating the contributions of different features to the model's forecasts, the SHAP values shown in Figure 3 allowed us to pinpoint the most valuable features concerning the mental health disorder prediction.

According to the interpretation, 40% of respondents who report having a family history of mental illness are far more likely to request therapy than those who do not. The model parameters have been adjusted through experiments with different sample values to strike a compromise between computation cost and explanation accuracy. The experiments

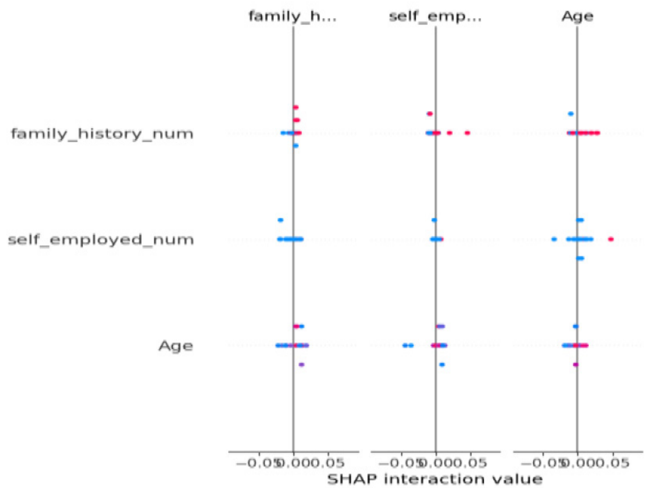

Figure 3

SHAP Variable Importance Plot - Global Interpretation

yield the following findings: (i) Explanations based on small numbers (like 50) are brief but vague. (ii) High levels (such as 1000) produce interpretations that are accurate but expensive to compute. A performance comparison of the suggested models with conventional ML models is shown in Table 1.

Table 1

Comparison of Conventional Machine Learning Models with LIME and SHAP

Model	Accuracy	Precision	Recall	F1-Score
LR	80.87%	78.9%	80.1%	79%
SVM	85.8%	82.4%	85.5%	84.2%
KNN	87.24%	84.4%	86%	87.2%
Multilayer Perceptron	91.94%	88.7%	90.5%	89.8%
LIME	97.6%	95.8%	97%	96.8%
SHAP	97.6%	95.8%	97%	96.8%

XAI models' ethical implications - There are more significant ethical and legal considerations as explainable AI (XAI) becomes more widely used in healthcare. Patient data privacy is a top priority because XAI techniques frequently need access to sensitive medical data in order to produce explanations. Techniques like federated learning and Differential privacy can ease perils by keeping patient data localized and anonymized during explanation generation.

5. Conclusion and Future Scope

Current research focuses on the notion that tech companies might start to foster a culture of empathy and understanding by offering their employees mental health benefits. Furthermore, it is not only excellent for the company to have happy and well-cared-for staff. Since the majority of them are young, age can also be a trigger because they are more likely to be receptive to receiving therapy. Interference at work is another major factor influencing employees' desire for therapy. In the proposed model, we were able to conclude that, in comparison to the "black box" nature of conventional machine learning models, the employment of XAI frameworks and toolkits, such as SHAP and LIME, can be extremely beneficial when interpreted in conjunction with other clinical considerations. Trends can be readily examined with partial dependency plots, opening the door for clinical diagnostics carried out in conjunction with explainable AI.

References

[1] V. Mehta, N. Batra, S. Goyal, A. Kaur, K. V. Dudekula, and G. J. Victor, "Machine learning based exploratory data analysis (EDA) and diagnosis of chronic kidney disease (CKD)," *EAI Endorsed Transactions on Pervasive Health & Technology,* vol. 10, no. 1 (2024).

[2] Puranam Revanth Kumar, B. Shilpa, Rajesh Kumar Jha "BrainTract: segmentation of white matter fiber tractography and analysis of structural connectivity using hybrid convolutional neural network", *Neuroscience*, Vol. 580, pp. 218-230 (2025).

[3] T. He and N. Lia, "Age-structured model with varied contact patterns and vaccination of COVID-19 in Liaoning Province, China," *Journal of Statistics and Management Systems,* vol. 27, no. 8, pp. 1525–1539 (2024).

[4] P. R. Kumar, R. K. Jha, P. A. Kumar, and B. D. Raju, "Improved neurological diagnoses and treatment strategies via automated human brain tissue segmentation from clinical magnetic resonance imaging," *Intelligent Medicine,* vol. 4, no. 3, pp. 161–169 (2024).

[5] B. Chen, L. Wang, B. Li, and W. Liu, "Work stress, mental health, and employee performance," *Frontiers in Psychology*, vol. 13 (Nov. 2022).

[6] P. R. Kumar, R. K. Jha, and P. A. Kumar, "BrainHyperintensities: Automatic segmentation of white matter hyperintensities in clinical brain MRI images using improved deep neural network," *The Journal of Supercomputing*, vol. 80, pp. 15545–15581 (2024).

[7] K. Kavita, R. Kulkarni, A. H. Prasad, B. Jeyakumar, M. Thaile, and S. Gore, "A novel optimization-based blockchain technology using health care data for enhancing security and privacy in the medical system," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 27, no. 8, pp. 2483–2494 (2024).

[8] P. R. Kumar, B. Shilpa, and R. K. Jha, "Brain disorders: Impact of mild SARS-CoV-2 may shrink several parts of the brain," *Neuroscience & Biobehavioral Reviews,* vol. 149, pp. 105150 (2023).

[9] M. Fabietti, G. Di Lazzaro, A. A. Steixner-Kumar, J. Modolo, E. Boccardi, and A. Schulze-Bonhage, "Artifact detection in chronically recorded local field potentials using long-short term memory neural network," in 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), *IEEE,* pp. 1–6 (Oct. 2020).

[10] P. R. Kumar, R. K. Jha, and A. Katti, "Brain tissues segmentation in neurosurgery: A systematic analysis for quantitative tractography approaches," *Acta Neurologica Belgica*, vol. 124, pp. 1–15 (2023).

[11] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy,* vol. 23, no. 1, pp. 18 (Dec. 2020).

[12] P. R. Kumar, B. Shilpa, R. K. Jha, and S. N. Mohanty, "A novel end-to-end approach for epileptic seizure classification from scalp EEG data using deep learning technique," *International Journal of Information Technology*, vol. 15, pp. 4223–4231 (2023).

[13] M. B. Khan and A. K. J. Saudagar, "Comparative analysis of medical knowledge and inferential capabilities of modern LLMs," *Journal of Statistics and Management Systems*, vol. 27, no. 8, pp. 1713–1722 (2024).

[14] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38 (Feb. 2019).

[15] P. R. Kumar, R. K. Jha, and P. A. Kumar, “Segmentation of White Matter Lesions in MRI Images Using Optimization-based Deep Neural Network,” in Proc. 4th Int. Conf. Image Process. Capsule Netw., pp. 253–267 (2023).

[16] K. N. Fitzgerald, R. R. Hodges, D. M. Hanes, M. Stack, D. Cheishvili, R. Szyf, and J. D. Bradley, “Potential reversal of epigenetic age using a diet and lifestyle intervention: A pilot randomized clinical trial,” Aging, vol. 13, no. 7, pp. 9419–9432 (Apr. 2021).

[17] P. R. Kumar, A. Katti, S. N. Mohanty, and S. N. Senapati, “A Deep Learning-based Approach for an Automated Brain Tumor Segmentation in MR Images,” *Pattern Recognit. Data Anal. Appl.*, vol. 888, pp. 87–97 (2022).

[18] L. J. Reece, C. McInerney, J. Blazek, D. Foley, B. Schmutz, R. Bellew, and J. Brown, “Reducing financial barriers through the implementation of voucher incentives to promote children’s participation in community sport in Australia,” *BMC Public Health,* vol. 20, no. 1, pp. 19 (Dec. 2020).

Received December, 2024