International Journal of Innovative Trends in Engineering Science and Technology
ijitest.org | E-ISSN: 3139-6887

Machine Learning-Based Employee Stress Prediction for Proactive Workplace Well-Being Management

[1]Mahendra Narla, [2]CH. Swapna Priya, , [3]K.Suseela, [4]T. Dansy Sri Abeesh
[1]Associate Professor, [2]Assistant Professor, [3]SGT Teacher, [4]Student
[1,4]Department of Artificial Intelligence and Data Science, [3]Department of School education, [4]Department of Computer Science Engineering
[1]G. Pullaiah College of Engineering & Technology, Kaurnool [2,4]Vignan's Institute of Information Technology, Visakhapatnam,
swapnachsp@gmail.com, dansysriabeesh@gmail.com, jackbenison12@gmail.com, narlamahendracai@gpcet.ac.in

Abstract -- Employee stress has become an important organizational concern due to evolving workplace environments, increasing professional responsibilities, workload variations, and changing work-life expectations. The widespread adoption of remote and hybrid working practices has further introduced challenges that can influence employees' psychological well-being and professional performance. Early identification of workplace stress is therefore essential for organizations to implement appropriate preventive measures and maintain a healthy working environment. This study proposes a machine learning-based framework for predicting employee stress levels using workplace-related and demographic attributes. Initially, data preprocessing and exploratory data analysis are performed to identify significant patterns, relationships, and variations among the input features. Subsequently, the XGBoost classification algorithm is employed to classify employees according to their stress levels. The predictive performance is evaluated using appropriate classification metrics to examine the effectiveness and reliability of the proposed approach. Furthermore, feature importance analysis is incorporated to identify the major factors contributing to employee stress. The findings highlight working hours, workload, age, and role ambiguity as important predictive factors associated with employee stress. The proposed framework can assist organizations and human resource management teams in identifying employees who may be at risk of workplace stress and adopting timely, data-driven interventions. This approach provides a foundation for developing proactive employee well-being strategies and promoting a more supportive and productive organizational environment.

Keywords: Employee Stress Prediction, Machine Learning, XGBoost Classifier, Workplace Well-Being, Predictive Analytics, Feature Importance, Human Resource Management, Stress Classification.

I. INTRODUCTION

Employee well-being is an essential factor influencing organizational productivity, workforce stability, and overall business performance. In modern working environments, employees frequently encounter professional challenges arising from increased workloads, extended working hours, demanding responsibilities, and difficulties in maintaining a healthy work-life balance. These factors can contribute to psychological pressure, reduced concentration, emotional exhaustion, and declining job satisfaction. The transition toward remote and hybrid working arrangements during and after the COVID-19 pandemic has further changed traditional workplace practices. Although these arrangements provide flexibility, they may also introduce challenges such as social isolation, communication gaps, blurred boundaries between personal and professional activities, and increased work-related pressure. Consequently, identifying employees who are potentially experiencing workplace stress has become an important requirement for organizations.

Traditional approaches to employee stress assessment generally depend on questionnaires, periodic surveys, personal feedback, and observations conducted by human resource professionals. Although these methods provide useful information about employee experiences, they may not consistently identify emerging stress conditions at an early stage. Furthermore, subjective responses, limited assessment frequency, and variations in individual perceptions can affect the effectiveness of conventional evaluation techniques. Recent developments in data analytics and machine learning have created opportunities to examine employee-related information and identify patterns associated with workplace stress. Machine learning algorithms can process multiple demographic and occupational attributes, discover relationships among variables, and generate predictive classifications. In particular, classification algorithms such as

DOI: 10.5281/zenodo.24591540
IJITEST-2026-034 | Volume 1, Issue 8, October 2026 | ISSN: 3139-6887
Creative Commons Attribution 4.0 International (CC BY 4.0)
Article Received: November 12, 2026 | Accepted: November 26, 2026

Decision Tree, Random Forest, Support Vector Machine, and Extreme Gradient Boosting (XGBoost) can support the identification of potential stress-related patterns in organizational datasets.

Motivated by these challenges, this research presents a machine learning-based framework for predicting employee stress levels and supporting proactive workplace well-being management. The proposed approach utilizes employee-related attributes to perform data preprocessing, exploratory analysis, and predictive classification using the XGBoost algorithm. In addition to stress prediction, feature importance analysis is conducted to examine the contribution of individual variables to the classification results. Particular attention is given to factors such as working hours, workload, age, and role ambiguity, which may provide useful insights into workplace stress patterns. By identifying employees who may require additional support, the proposed framework can assist human resource departments in planning preventive interventions, improving workload distribution, and promoting healthier working conditions. The primary objective is to establish a data-driven approach that supports early stress-risk identification rather than relying exclusively on reactive employee assessments.

II. Related Work

Cooper and Marshall [1] examined the major sources of occupational stress and discussed their relationship with employees' physical and psychological health. Their review highlighted the importance of organizational and job-related conditions in understanding workplace stress. Similarly, Karasek [2] introduced the Job Demand–Control model, which explains how the combination of high job demands and limited decision-making authority can contribute to mental strain among employees. This theoretical perspective establishes the importance of considering workload, job responsibilities, and working conditions when evaluating occupational stress.

Bakker and Demerouti [3] presented the Job Demands–Resources (JD-R) model, which explains the relationship between job demands, available workplace resources, and employee well-being. According to this model, excessive job demands can increase the possibility of exhaustion, whereas appropriate organizational resources can help employees manage professional pressures. These theoretical contributions provide a foundation for selecting workplace-related attributes for employee stress prediction.

Machine learning techniques provide computational approaches for identifying complex patterns in structured datasets and predicting outcomes based on historical observations. Classification algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and ensemble-based methods are widely used in predictive analytics. These techniques can be applied to employee-related datasets to investigate associations between demographic characteristics, occupational conditions, and stress categories. Unlike conventional statistical approaches that may require predefined relationships between variables, tree-based ensemble methods can capture nonlinear interactions among input features. Therefore, machine learning offers a promising approach for developing employee stress prediction systems. However, model selection and performance should be validated using appropriate evaluation metrics and independent test data.

Chen and Guestrin [4] introduced Extreme Gradient Boosting (XGBoost), an optimized tree-boosting framework designed to improve computational efficiency and predictive performance. Their approach incorporates regularization, sparsity-aware learning, and parallel processing to handle structured datasets effectively. These characteristics make XGBoost a suitable candidate for classification problems involving multiple employee-related attributes. Nevertheless, predictive accuracy alone may not provide sufficient information about the factors influencing individual predictions. To address this limitation, Lundberg and Lee [5] proposed SHapley Additive exPlanations (SHAP), a unified framework for interpreting machine learning predictions through feature contribution values. SHAP can help explain how individual attributes influence model outputs and provide greater transparency in predictive systems. Based on these developments, the present research investigates an XGBoost-based framework for employee stress prediction and examines the importance of workplace and demographic attributes to support proactive employee well-being management.

III. Proposed Methodology

The proposed methodology begins with the collection of structured employee datasets containing demographic information, occupational characteristics, and behavioral attributes. The dataset includes relevant parameters such as working hours, workload, job responsibilities, and role ambiguity, which are considered important for analyzing employee stress patterns. These collected records serve as the primary input for developing the employee stress prediction model.

After data collection, exploratory data analysis (EDA) is performed to understand the structure and characteristics of the dataset. Statistical analysis and visualization techniques are applied to examine feature distributions, identify relationships among variables, detect unusual observations, and understand the representation of stressed and non-stressed employees. This analysis provides useful insights into the dataset and helps determine the relevant attributes for the prediction process.

Subsequently, data preprocessing and feature transformation techniques are applied to prepare the dataset for machine learning. Missing values are handled using suitable imputation methods, while categorical attributes are converted into numerical representations through label encoding. Relevant features are selected based on their significance in employee stress prediction. These preprocessing operations improve data consistency and ensure that the input variables are in an appropriate format for model development.

To address possible class imbalance in the dataset, suitable data resampling techniques are applied to maintain a balanced representation of different stress categories. This step helps reduce the possibility of the classifier favoring the majority class and improves its ability to recognize patterns across all target categories. The processed dataset is then divided into training and testing subsets, with 85% of the data allocated for model training and the remaining 15% reserved for testing and performance assessment.

The training dataset is provided to the XGBoost (Extreme Gradient Boosting) classifier to develop the employee stress prediction model. XGBoost uses an ensemble of decision trees, where successive trees are constructed to minimize the errors generated by previous trees. Its gradient boosting mechanism and regularization capabilities help improve predictive performance while controlling model complexity. The trained model learns relationships between employee-related attributes and their corresponding stress categories to generate predictions for previously unseen records.

Finally, the trained model is evaluated using the independent testing dataset to examine its ability to classify employee stress levels accurately. Standard evaluation measures, including accuracy, precision, recall, F1-score, True Positive Rate, and False Positive Rate, are considered to assess classification performance. A confusion matrix is also used to analyze correct and incorrect predictions across the target classes. The evaluation results help determine the effectiveness of the proposed XGBoost-based framework in identifying potential employee stress conditions and supporting early intervention strategies for improved workplace well-being.

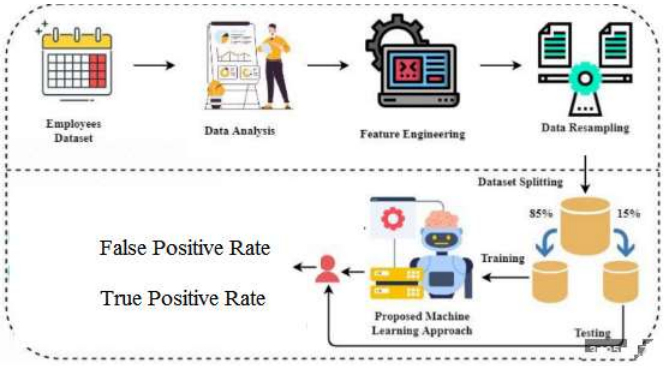

Figure 1: Proposed Machine Learning Pipeline for Employee Stress Prediction and Evaluation.

IV. Results and Discussions

The stress distribution analysis illustrates the number of employees classified into stressed and non-stressed categories based on the available dataset. The horizontal axis represents the two target classes, where 0 indicates non-stressed employees and 1 represents stressed employees, while the vertical axis displays their corresponding record counts. The graphical representation helps examine the proportion of employees belonging to each category. The observed distribution indicates that the non-stressed group contains a higher number of records than the stressed group. This difference suggests the presence of class imbalance in the dataset. Identifying this variation during exploratory data analysis is important for selecting suitable balancing techniques and developing a reliable classification model.

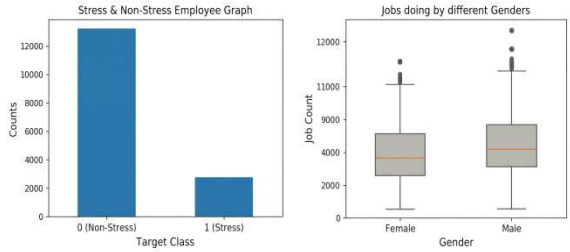

Figure 2: Exploratory Data Analysis: Target Class Distribution (Stress vs. Non-Stress) and Job Count by Gender

The gender-based job distribution analysis examines how job-related responsibilities are represented among male and female employees. The horizontal axis categorizes employees according to gender, whereas the vertical axis indicates the corresponding job count values. The box plot presents the distribution of job-related observations through median values, quartiles, and variations within each gender group. These statistical measures help understand differences in the distribution of job roles across the workforce. The comparison provides insights into demographic characteristics and possible variations in occupational participation. Such exploratory analysis supports a better understanding of employee attributes before applying feature selection and machine learning techniques.

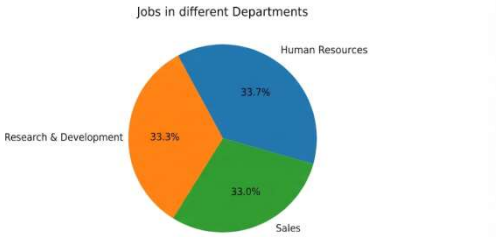

Figure 3: Employee Distribution Across Organizational Departments

This pie chart illustrates the percentage distribution of job roles across three major organizational departments. Specifically, it highlights that Human Resources accounts for 33.7%, Research and Development constitutes 33.3%, and Sales represents 33.0% of the total roles. This balanced proportional breakdown provides a clear overview of workforce allocation across different corporate segments.

The box plot illustrates the relationship between job involvement levels and individual contributor status among

employees. The horizontal axis represents the employee categories based on their individual contributor classification, while the vertical axis indicates their corresponding job involvement scores. Different colored box plots distinguish between stressed and non-stressed employees, enabling a comparative assessment of their engagement levels. The median values and interquartile ranges provide insights into the variation and distribution of job involvement across the groups. This visualization helps examine whether individual contributor responsibilities are associated with differences in employee engagement and stress patterns. The findings can support the identification of relevant workplace attributes for further analysis and employee stress prediction using machine learning techniques.

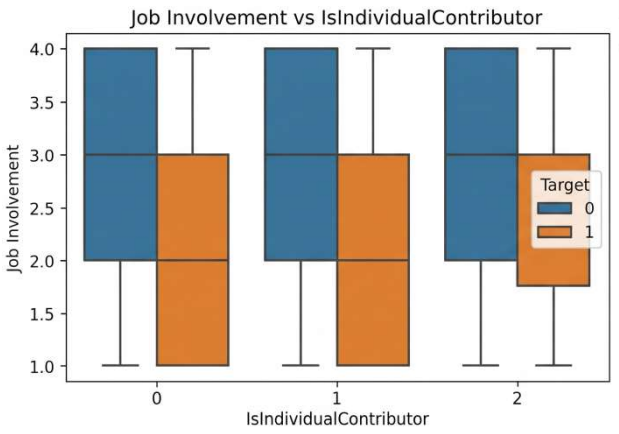

Figure 4: Job Involvement Distribution Across Individual Contributor Categories by Stress Class

The dataset correlation heatmap provides a comprehensive visual representation of the linear relationships established among various structural and behavioral attributes within the employee records. Within this analytical matrix, individual features are mapped along both axes to compute pairwise correlation coefficients, highlighting how strongly different variables move together. Darker color gradients typically signify weaker or negligible statistical association, whereas intensified highlights point toward stronger multicollinearity or dependencies among specific features. By examining these correlation values, researchers can effectively identify which input predictors exert the most pronounced statistical influence on target outcomes like employee stress. This critical diagnostic stage ensures that redundant or uninformative columns are properly addressed prior to advanced feature scaling and model training. Ultimately, analyzing these feature-to-feature interactions safeguards the machine learning pipeline against unnecessary noise and enhances overall predictive stability.

The XGBoost confusion matrix evaluates the classification performance of the model by comparing predicted employee stress outcomes against actual ground-truth labels on the test set. Specifically, the matrix demonstrates 1,292 True Negatives (correctly predicted non-stress cases) and 112 True Positives (correctly predicted stress cases). However, it also records 173 False Negatives (missed stress cases) and 23 False Positives (incorrectly flagged stress cases), indicating that while the model excels at identifying normal conditions, it has a higher tendency to overlook actual instances of stress.

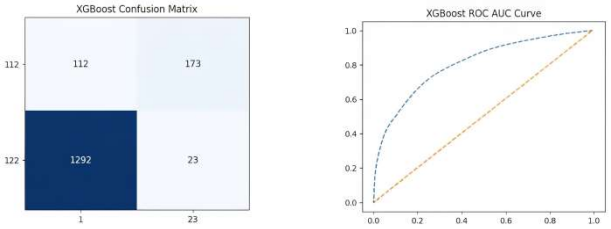

Figure 5: XGBoost Classification Performance: Confusion Matrix and ROC-AUC Curve

Complementing this, the ROC-AUC curve visualizes the model's overall discriminatory power by plotting the True Positive Rate against the False Positive Rate across various classification thresholds. The curve achieves an AUC score of approximately 0.797, rising steeply initially to indicate strong performance at low false-positive rates. Nevertheless, the curve flattens before reaching the top-right corner, confirming that the classifier faces challenges in capturing all positive stress cases without increasing false positives, pointing toward potential class imbalance.

V. Conclusion and Future Scope

This study presents a machine learning-based framework for predicting employee stress levels using demographic and workplace-related attributes. The XGBoost classifier is employed to identify potential stress conditions based on the available dataset. Data preprocessing and exploratory analysis help discover meaningful patterns among employee-related features. Feature importance analysis identifies working hours, workload, age, and role ambiguity as important predictive attributes. The findings demonstrate the potential of machine learning techniques in identifying workplace stress patterns and supporting early risk assessment. The proposed framework can assist organizations and human resource professionals in planning appropriate preventive measures and improving employee well-being. In future, advanced algorithms such as Random Forest, LightGBM, CatBoost, and deep learning models can be explored to enhance prediction performance. Explainable AI techniques, including SHAP and LIME, can provide better insights into individual predictions. Additional factors such as sleep quality, job satisfaction, work-life balance, and organizational support can be incorporated to improve prediction reliability. Furthermore, the framework can be extended into a real-time, web-based stress monitoring system while ensuring employee data privacy and confidentiality. These developments can support the creation of an intelligent and effective workplace stress management system.

REFERENCES

[1] M. J. Hasan, J. Sultana, S. Ahmed, and S. Momen, "Early detection of occupational stress: Enhancing workplace safety with machine learning and large language models," *PLOS ONE*, vol. 20, no. 6, Art. no. e0323265, 2025. DOI: https://doi.org/10.1371/journal.pone.0323265

[2] S. Muñoz, C. Á. Iglesias, O. Mayora, and V. Osmani, "Prediction of stress levels in the workplace using surrounding stress," *Information Processing & Management*, vol. 59, no. 6, Art. no. 103064, 2022.

DOI: https://doi.org/10.1016/j.ipm.2022.103064

[3] Y. Zhang and E. Qi, "Happy work: Improving enterprise human resource management by predicting workers' stress using deep learning," *PLOS ONE*, vol. 17, no. 4, Art. no. e0266373, 2022.

DOI: https://doi.org/10.1371/journal.pone.0266373

[4] "Semi-supervised and ensemble learning to predict work-related stress," *Journal of Intelligent Information Systems*, 2024.

DOI: https://doi.org/10.1007/s10844-023-00806-z

[5] "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *Journal of Biomedical Informatics*, vol. 139, Art. no. 104299, 2023.

DOI: https://doi.org/10.1016/j.jbi.2023.104299

[6] F. Rodrigues and H. Correia, "Semi-supervised and ensemble learning to predict work-related stress," *Journal of Intelligent Information Systems*, vol. 62, pp. 77–90, 2024.

DOI: https://doi.org/10.1007/s10844-023-00806-z

[7] M. Naegelin, R. P. Weibel, J. I. Kerr, V. R. Schinazi, R. La Marca, F. von Wangenheim, C. Hoelscher, and A. Ferrario, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *Journal of Biomedical Informatics*, vol. 139, Art. no. 104299, 2023.

DOI: https://doi.org/10.1016/j.jbi.2023.104299

[8] J. Kallio, E. Vildjiounaite, J. Tervonen, and M. Bordallo López, "A survey on sensor-based techniques for continuous stress monitoring in knowledge work environments," *ACM Transactions on Computing for Healthcare*, vol. 6, no. 3, Art. no. 36, 2025.

DOI: https://doi.org/10.1145/3720551

[9] M. Koch and M. Lodefalk, "Artificial intelligence and worker stress: Evidence from Germany," *Digital Society*, vol. 4, Art. no. 5, 2025.

DOI: https://doi.org/10.1007/s44206-025-00160-3

[10] L. F. Pareja Bernal, S. Ben Souissi, and C. Golz, "Machine learning-based detection of workplace stress using wearable and multimodal data: A systematic literature review," *Frontiers in Artificial Intelligence*, vol. 9, Art. no. 1837195, 2026.

DOI: https://doi.org/10.3389/frai.2026.1837195

Funding Declaration

The authors declare that no funds, grants, or other forms of financial support were received from any organization or institution for the conduct of this research.