International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Data-Driven Stress Assessment for Professionals

Trushti D Patel[1], Mrs. Liyansi Patel[2]

[1]Computer Engineering Department, Gujarat Technological University, Chandkheda, Ahmedabad

[2]Computer Engineering Department, KIRC, Kalol

[1]pateltrushti222@gmail.com

[2]Ptelliyansi9743@gmail.com

Abstract

Job stress has become one of the major challenges of professional life in the modern age and it impacts the mental well-being, job output and general living. This paper, Data-Driven Stress Assessment for Professionals, aims to look at different aspects of stress including lifestyle habits, working patterns and routines with the use of machine learning to determine how the above factors affect stress. It contains the data of 50000 subjects with 17 variables ranging over demographics (age, gender, country and occupation), work-life balance measures (working hours, sleep duration and physical activity), and lifestyle-based choices (diet quality, smoking, alcohol consumption and social media consumption). Mental health data, such as stress reported by the patients themselves, the history of consultation and the medication as well as their severity further enrich the data and increase clinical value of the whole data set.

The research design includes the process of cleaning and preparing both numeric and categorical features and creating new features and use supervised machine learning approaches to predict the degree of stress. The performance of such models as Random Forest, XGBoost, and Neural networks will be tested in terms of their accuracy, and statistical analysis will be introduced to consider the most important factors linked to occupational stress. The research goes further to compare the variation in stress by the different age groups, gender, regions with the aim of clarifying the stress pattern according to the different demographics.

The desired outcomes are the availability of a consistent and decipherable model that can suggest the levels of stress with a high degree of accuracy and identify dominant risk factors, i.e., a long working schedule, inadequate sleep, the absence of physical activity, and negative lifestyle choices. This knowledge can guide employers, policymakers, and mental health professionals to come up with more effective interventions, workplace programs on wellness, and early warning systems in regards to individuals considered at risk of developing chronic stress. This project will contribute to a growing body of research in the burgeoning field of computational mental health as well as to efforts to provide evidence-based support to the well-being of professionals.

Keywords: **Occupational Stress, Machine Learning, Artificial Neural Networks (ANN), Stress Prediction, Computational Mental Health**

I. INTRODUCTION

Occupational stress has become a pressing issue of the contemporary work environment in the past few years, impacting greatly on the productivity of employees, their mental health, and the overall performance of the organization in general. The changing work environments, the growing workloads, and the changing lifestyle patterns have been some of the contributing factors to the high stress levels among the working professionals. Conventional techniques of stress measurement which include questionnaires and interviews are generally subjective, time consuming, and also cannot be used to measure intricate relations between various factors influencing the measurement.

As data-driven technologies progressed, machine learning (ML) and deep learning (DL) methods demonstrated huge potential in the analysis of massive data collections and the discovery of previously unseen patterns in behavioural and occupation data. Such computing methods allow more precise, scalable, and objective prediction models of stress as opposed to the traditional methods. Nevertheless, any research in the field is most typically limited to a small sample, small number of features, or specific occupations, which limits their applicability and utility.

To overcome these weaknesses, this study puts forward an evidence based stress prediction model by collecting a rich dataset of 50,000 working professionals incorporating all the various features including lifestyle habits, working environment conditions, and demographics. Various machine learning models are used to test and compare the performance of the models developed and they are; Logistic Regression, Random Forest, and Artificial Neural Networks (ANN).

The findings indicate that stress is a multifactorial phenomenon being influenced by many factors that are interrelated as working hours, sleep patterns, work-life balance, and lifestyle behaviors. The ANN model is one of the models tested and it is known to be effective in identifying nonlinear relationships and has a high prediction accuracy hence it is highly applicable in the stress classification activities.

The work has an addition with the area of computational mental health as it offers a scalable and data-driven model of stress analysis. The suggested framework could help organizations adopt proactive stress monitoring systems that will allow the prompt recognition of the high-risk individuals and the application of specific interventions to enhance the well-being of employees and the efficiency of the organization.

A. Scope

The target audience used in this research is the workforce of various dimensions and sectors that include IT sector, financial sector, healthcare sector, and the education sector. The data is self-reported data on lifestyle and work patterns and mental health metrics. The research does not tend to give clinical diagnosis but rather give estimates on levels of stress and expose contributory factors.

There are several countries involved in the geographic scope where cross-cultural experiences can be made. Nevertheless, since the data is of the survey nature, it is subject to the integrity of the responses given by the participants. It is not a comprehensive investigation into mental disorders beyond the identification of their stress types but the introduction of machine learning models that just presuppose its research but are not ready to be applied in practice, at least immediately.

B. Research Problem

Despite widespread recognition of workplace stress as a major threat to individual and organizational performance, existing measurement methods remain inconsistent, subjective, and difficult to scale, often relying on biased self-reports that may underrepresent or overstate actual stress levels. Much of the current

research is limited to small, localized samples, restricting its applicability across diverse populations, especially in today's global and digitally evolving work environments where new stressors—such as blurred work-life boundaries and increased digital exposure—are not adequately captured by traditional tools. This lack of accurate, scalable assessment methods hinders early detection and leads to reactive rather than proactive interventions, resulting in reduced productivity, financial losses, and declining employee well-being. To address this gap, there is a strong need for data-driven and algorithm-based approaches; therefore, this study utilizes a large dataset of 50,000 individuals with multiple features related to demographics, lifestyle, and mental health, applying machine learning techniques such as Random Forest, XGBoost, and Neural Networks to develop scalable, objective, and predictive models for stress detection and management.

C. *Research Problem*

1. Which lifestyle and occupational factors are most predictive of stress levels?
2. Can machine learning models accurately estimate stress in large, heterogeneous professional populations?
3. How do demographic differences (age, gender, country) influence stress patterns?
4. What model provides the best balance between prediction accuracy and interpretability?

D. *Objectives*

- To build a robust machine learning framework for stress level classification
- To identify key determinants of occupational stress using data-driven techniques
- To study demographic and occupational inequalities in stress level
- To evaluate and compare multiple machine learning models for stress prediction
- To provide actionable insights for workplace wellness and mental health policy
- To explore correlations between mental health history and stress patterns

II. Literature Review

A. *Problem Statement*

Stress in the workplace has become a rising issue in the constantly busy modern workplaces and it affects productivity, mental health, and life quality in general. Compared to the conventional ways of measuring stress (surveys, interviews, or the clinical assessment), which tend to be subjective, time-consuming (they require many resources) and can only involve a small group of individuals, the adaptation of HRV can measure stress levels on a larger scale, in a more objective manner, and can be conducted remotely. Such methods fall short in drawing scalable, evidence-based conclusions about the role lifestyle factors, work habits, and demographic characteristics and their contributions to stress in different professionals.

Current academic studies present correlations between overwork, inadequate sleep, poor lifestyle behavior and mental disorders and yet there exist no quantitative frameworks that can help achieve error-free predictive models and further substantiate such relationships. Moreover, studies are mostly lacking in terms of including the cross-demographical differences (such as age, gender, occupation, etc.) and the

influence of previous mental health history. Indeed, more advanced machine learning techniques have the promise of overcoming these challenges, but are seldom used or systematically compared.
This necessitates a sound, data-based solution to not only predict stress levels properly but to also be able to determine the essential influencing factors. This type of approach would allow the identification of individuals at high-risk as early as possible, and could generate actionable intelligence about employees on the organizational level and help with the design of workplace wellness initiatives and mental health policies.

B. *Research Gaps*

Q1. Why is there no reliable large-scale model for stress estimation among working professionals?
A1. Existing studies are often limited to small datasets or narrow professions. This creates a gap in building machine learning models that can generalize across diverse occupations and demographics.
Q2. Which lifestyle and occupational factors most strongly predict stress levels?
A2. Current research often analyzes isolated variables rather than integrating multiple behavioral, demographic, and occupational factors to understand their combined effect on stress.
Q3. How do stress patterns vary across demographic groups and job sectors?
A3. Few studies explore stress disparities by age, gender, or profession, limiting the ability to design targeted wellness strategies.
Q4. Which machine learning techniques are best suited for accurate and interpretable stress prediction?
A4. Research rarely compares multiple algorithms or addresses the "black box" nature of advanced models, leaving uncertainty about which approach balances accuracy with transparency.
Q5. How can research outcomes translate into actionable workplace interventions?
A5. Even when stress factors are identified, most studies fail to provide clear recommendations for employers, mental health professionals, or policymakers.
Q6. What is the relationship between mental health history and current stress levels?
A6. Prior consultation history, medication use, or existing mental health conditions are often ignored in stress estimation models, leaving a critical gap in understanding risk patterns.

C. *Research Paper*

The study by (Chandrasekaran et al., 2025) investigated work-related stress among women in public and private sectors in Tamil Nadu, India, using a cross-sectional survey of 200 participants across multiple professions. Statistical analyses revealed that evaluation by superiors and negative feedback were the most significant stressors, followed by long working hours, lack of emergency leave, and pay disparities, with public sector employees experiencing authority pressure and private sector employees facing workload stress. Coping strategies such as spending time with children and spiritual practices were commonly adopted, while yoga and meditation were least utilized. The study is limited by its regional focus, self-reported data, and lack of longitudinal analysis.

The research by (Mahajan et al., 2025) examined lifestyle-related determinants among 208 IT professionals in Pune using tools such as PSS and IPAQ, identifying high levels of physical inactivity, moderate stress, and unhealthy lifestyle behaviors including poor diet and excessive screen time. A significant proportion of participants were overweight or obese, indicating increased risk of non-communicable diseases, though no gender differences were observed. Limitations include self-reported data, small sample size, and cross-sectional design.

Wu et al. (2025) explored the relationship between occupational stress and mental health outcomes among 657 Chinese nurses using ERI, HADS, and AIS scales, finding that high stress levels significantly increased anxiety, depression, and insomnia. Nurses with effort-reward imbalance and low social support were particularly vulnerable, especially younger staff with heavy workloads. The study is limited by its cross-sectional design and reliance on self-reported measures.

Alharbe (2025) applied the fuzzy analytic hierarchy process (F-AHP) to analyze factors contributing to stress, anxiety, and depression, identifying poor sleep, chronic illness, financial strain, and lack of social support as major contributors. Among interventions, adequate sleep and physical activity were ranked most effective. While F-AHP provided structured prioritization, the study relied on expert judgment rather than empirical data, limiting generalizability.

Filippis and Foysal (2024) conducted a machine learning-based analysis on 1,100 student records using Random Forest models, achieving approximately 0.88 accuracy in predicting stress. The study found strong correlations between anxiety and academic performance and between sleep quality and depression, while bullying and peer pressure emerged as key predictors. Limitations include self-reported data and lack of causal inference.

Breuer-Asher et al. (2024) investigated the relationship between smartphone usage and academic performance among 400 university students in Vietnam, finding that excessive non-academic smartphone use negatively correlated with GPA, while moderate academic use showed positive effects. Students averaged 6.5 hours of daily usage, primarily for entertainment. The study is limited by self-reported data and cross-sectional design.

Al-Alim et al. (2024) proposed machine learning models for stress detection using physiological signals such as heart rate variability, electrodermal activity, and respiration rate, demonstrating that ensemble and deep learning methods achieved higher accuracy than traditional models. The study emphasizes multimodal data integration but notes challenges related to signal noise, dataset size, and real-world implementation.

Mitra and Sharma (2025) developed a deep learning-based stress detection system using sentiment analysis of textual data, employing CNN models with NLP preprocessing techniques. The model showed improved accuracy compared to traditional classifiers, confirming the effectiveness of linguistic features in detecting stress, though limitations include domain dependency and lack of multimodal integration.

Lazarou and Exarchos (2024) reviewed real-time stress prediction models using wearable devices and physiological data, highlighting signals such as heart rate variability and electrodermal activity combined with machine learning techniques. While accuracy is high in controlled settings, performance declines in real-world environments due to variability and noise. The study emphasizes personalization and context-aware modeling.

Dhanalakshmi and Dev (2024) analyzed workplace stress among 362 software professionals in Chennai, finding high stress prevalence driven by work-family imbalance, workload, and job insecurity. While moderate stress can enhance productivity, excessive stress negatively impacts well-being and performance. The study is limited by its regional scope and reliance on self-reported data.

Trivedi et al. (2024) examined work stress among 356 IT professionals during the COVID-19 pandemic using the TAWS-16 tool, reporting that stress was associated with deadlines, long hours, and work-life imbalance. Higher stress levels were observed among women and mid-career employees, though results lacked statistical significance. Limitations include self-reported data and restricted geographic scope.

Matsumoto et al. (2023) reviewed stress detection using wearable sensors, identifying physiological indicators such as heart rate variability and electrodermal activity as reliable measures. Multimodal approaches improved accuracy, though challenges include data variability, device limitations, and privacy concerns. The study emphasizes the need for real-world validation and large-scale deployment.

Masri et al. (2024) reviewed workplace stress assessment methods, comparing subjective self-reports with objective physiological measures, and highlighting the benefits of combining both approaches for accurate stress detection. While wearable technologies and AI offer promising solutions, challenges remain in standardization, scalability, and data privacy.

Alruily (2023) proposed a hybrid deep learning model combining CNN-LSTM with optimization algorithms for stress detection, achieving high accuracy of 99.8%. The model integrates physiological and sentiment data, outperforming traditional classifiers, though limitations include high computational complexity and dataset bias.

Shahapur et al. (2024) examined stress among IT professionals in India, identifying workload, job insecurity, and poor work-life balance as major contributors. Stress negatively affected productivity and well-being, while coping strategies included recreational activities and social support. The study is limited by self-reported data and lack of longitudinal analysis.

Haque et al. (2024) reviewed AI-based stress prediction using heart rate variability data, finding that deep learning models outperform traditional methods, while multimodal approaches improve accuracy. Challenges include lack of standardized datasets, small sample sizes, and limited real-world validation.

Hosseini et al. (2022) provided a dataset on occupational stress among 350 academics in Nigerian universities, identifying workload, role ambiguity, and work-life balance as key stressors. Younger faculty reported higher stress levels. The study is descriptive and limited by self-reported data and geographic scope.

Pabreja et al. (2021) applied machine learning models to predict stress levels among Indian professionals, finding that Random Forest and SVM achieved the highest accuracy. Key stressors included long working hours, poor sleep, and lack of exercise. Limitations include small dataset size and lack of longitudinal tracking.

Patil et al. (2021) studied stress among 105 critical care healthcare professionals, finding that stress levels were influenced by gender and sleep disturbances, with many participants reporting fatigue and anxiety. The study is limited by its small sample size and single-hospital setting.

Jukic et al. (2020) evaluated a mobile health-based stress management program among employees, reporting improved wellbeing, reduced stress, and better physiological markers such as cortisol levels. Despite promising results, the study is limited by a small sample size and lack of a control group.

III. METHODOLOGY

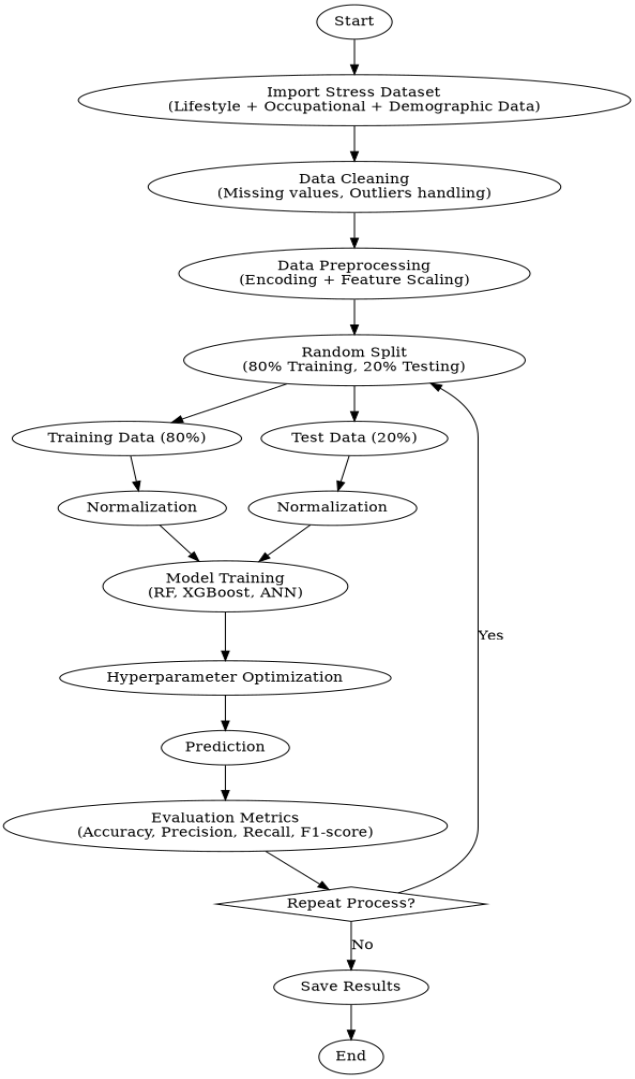


Fig.1 Proposed Flow of the Research

The proposed work starts with the importation of a detailed data of 50,000 working professionals with demographic, lifestyle, occupational, and mental health characteristics. The data has been obtained through the portal of NHANES 2021-2023 which is a well-regarded and popular type of data released by Centers of Disease Control and Prevention (CDC) and is characterized by its standardised data collection and structure that is nationally representative. It also covers a wide range of variables, which include age, gender, income, sleep habits, exercise, health issues, and laboratory findings, which make it very appropriate in the analysis of intricate relationships. This is followed by preprocessing to clean and standardize the data such as the treatment of missing values, normalizing numeric features, encoding categorical variables such as the use of Label Encoding in case of the preparation of the data to the machine learning models.

Following preprocessing, the data is divided into two sets, the training and the validation set in order to have appropriate model generalization. To develop the model, an EfficientNetB0-based architecture with tailor-made layers is applied because of the balance between its performance and computational efficiency. The model is gathered using the right optimizers, loss functions, and the evaluation measures of accuracy,

precision, and recall. To improve training, EarlyStopping is used to avoid overfitting and AUTOTUNE is used to optimise resources so as to maintain a stable and an optimised learning.

Lastly, several performance measures are applied to the trained model, such as accuracy, precision, recall, F1-score and confusion matrix, to give a multifaceted analysis of the predictive performance of the trained model. Such organized pipeline, i.e., data preparation-model evaluation, leads to a powerful and scalable system of predicting stress levels among working professionals. The framework can be further applied to the real world use like wellness program at the workplace, early stress detection systems and customized health recommendations.

A. Parameter Details

The psychological health and lifestyle data utilized in the present paper is a very important factor in the comprehension of the linkage between professional life, everyday habits, and stress. It includes 50,000 records and 17 distinct variables and will provide a rich understanding of the way the variables of demographics, occupation, and behavior interact with each other to determine stress levels among working professionals. Demographic characteristics like age, gender, occupation, and country are also relevant backgrounds and it is possible to compare the cross-sectional work-life periods, working conditions, and cultural backgrounds. The variables related to mental health, such as the severity of the condition, history of consultation, and medication use, contribute to the clinical richness, and the stress level is the main target one that will determine the predictive modeling.

Moreover, such lifestyle and behavioral aspects as sleep duration, working hours, physical activity, and social media use throw light on the influence of everyday habits on mental health. Such health related factors as the quality of the diet, smoking and alcohol consumption are additional attributes that strengthen the data set because they represent both adaptive and maladaptive coping. Combining these various dimensions, the data transcends the simplistic statistical representation, and the actual behavioral patterns of the real-life and allow the creation of more correct predictive models and helps to develop more individual-focused and based on the data approaches to the management of stress and healthier working conditions.

IV. IMPLEMENTATION

A. Analysis and Exploration of Data

The Corporate Stress Data employed in the study is comprised of about 50,000 employees records that have 30 variables of demographic, occupation, workplace, and psychological variables. It will focus on investigating the correlation of the work environment and mental health of an employee, and the main goal is to predict burnout symptoms with the help of machine learning. Some of its most prominent attributes are demographic information (age, gender), employment-related factors (position, experience, salary), working environment (working hours, commuting, sleep, work pressure, work life balance, managerial support, job satisfaction), and psychological elements (job pressure, work life balance, managerial support and job satisfaction). Also, such variables as training opportunities, discrimination, and gender bias are valuable understanding of organizational culture and fairness. Burnout symptoms (Yes/No) is the target variable which allows the model to reveal the employees at risk. On the whole, the given data can be used to create predictive models to assist organizations in taking active steps to alleviate stress, offer better workplace policies, and improve employee welfare.

Data Visualization

Fig 2: Relationship between working hours per week and stress levels for female employees.

The visualization demonstrates easily that there is a correlation between the number of hours working per week and the degree of stress among female employees with an average of 4.7 to 5.4 on stress levels influenced by working between 35 and 90 hours. Although the rise in stress is not sharp but is steady, it shows that the longer the working hours, the more the stress levels at any given time will always be high. The fairly constant trend however, is also indicative of the fact that work load is not the only factor that contributes to stress but also an array of other factors including work environment and support systems. On balance, the prolonged working hours correlate with the mental exhaustion, the decrease in the work-life balance, and the elevation of the burnout probability.

Relationship Between Sleep Hours and Stress Level

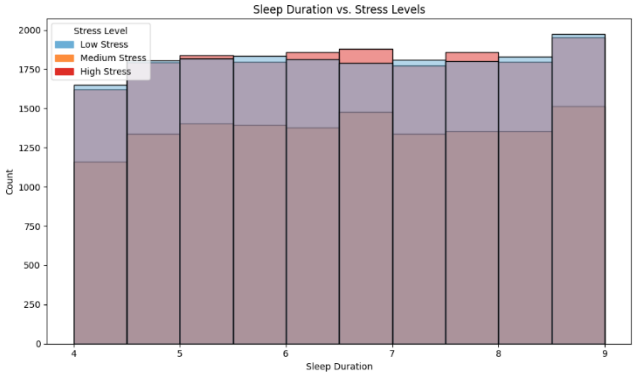


Fig 3: Relationship Between Sleep Hours and Stress Level

The above graph depicts that there is an inverse relationship between the length of time one is sleep deprived and the stress level of the employee so that sleep deprived employees (approximately 4-6 hours sleep) are more prone to high stress levels; on the other hand, well balanced sleep deprived employees (7-9 hours sleep) are less prone to stress. Despite the certain overlap of stress types, the general tendency is that sleep is associated with improved mental health and less stress. Sleep is effective in increasing emotional resilience, concentration, and the capacity to deal with work-related pressure. On the whole, the visualization explains that sleep is one of the main means of minimizing stress and avoiding burnout among workers.

Uniformity and Distribution of the Dataset

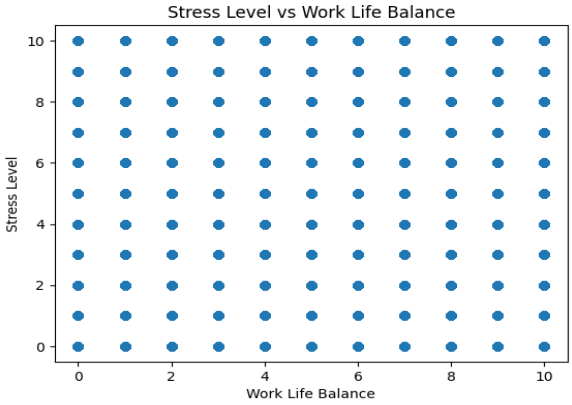


Fig 4: Uniformity and Distribution between Stress Level and Work Life Balance

An exploratory data analysis indicated that the dataset is relatively homogeneous i.e. the values are distributed uniformly across various values showing no major clustering or extreme outliers. This balanced distribution is observed through such characteristics as working hours (35-90 hours), sleep (4-9 hours), and stress (4.6-5.4) levels, which show that the variety of employees but at the same time realistic patterns can be observed. This kind of uniformity makes machine learning models able to generalize on a large variety of situations without committing to certain values. All in all, the widely-spread information facilitates the stability of the model, enhances the generalization, and contributes to the creation of precise predictive models to detect stress and burnout.

Insights from Exploratory Analysis

The exploratory data analysis helps to identify important factors contributing to employee burnout as the workplace conditions and lifestyle habits seem to be important contributors to the level of stress. The longer the working hours, the more stress and fatigue are observed, and proper sleep allows to sustain the stress at the same level and to feel better about the whole situation. The results have also pointed out that burnout is not brought about by any particular factor but a combination of various variables like work pressure, work-life balance, managerial support and personal habits. Moreover, the evenness of the distribution of data will increase the reliability of machine learning models and contribute to the use of neural networks to predict burnout correctly and assist organizations to adopt effective strategies to manage stress.

B. Data Preprocessing

Removing Unnecessary Columns

In the process of data preprocessing, some of the irrelevant columns are eliminated, including the ID field because it would not make a difference in the prediction of employee stress or burnout. The ID column is simply a unique identifier that does not give any meaningful information to be recognized in a pattern and its inclusion might add noise or false patterns in the training. Thus, it is discarded with the help of the command df = df.drop("ID" axis=1) so that the dataset includes only the features that are important. This enhances model efficiency, accuracy, and overall learning since emphasis is laid on variables that have only a single impact on predicting burnout.

Handling Categorical Variables

Categorical features such as Gender, Job Role, Department, and Location contain qualitative data and must be converted into numerical form for machine learning models to process them effectively. This is achieved using One-Hot Encoding, where each category is transformed into a separate binary column indicating its presence (1) or absence (0). For example, the Gender feature is converted into multiple columns like Male, Female, and Non-Binary, each represented with binary values instead of text. This transformation enables the neural network to interpret categorical data and learn meaningful patterns during training.

Example representation:

Gender	Male	Female	Non-Binary
Male	1	0	0
Female	0	1	0

Such change enables the machine learning model to read categorical information in a mathematical form without any implication of any numerical correlation between categories.

To implement the encoding process in Python the following command is used:

df = df.getdummies(dropfirst=True)

One-Hot Encoding is used to automatically convert all categorical variables into binary columns, making the dataset suitable for machine learning models. The parameter drop_first=True removes one category from each feature to reduce redundancy and prevent multicollinearity. This approach ensures that categorical data is represented without implying any false order, which could occur with simple numerical encoding. Overall, One-Hot Encoding transforms the dataset into a fully numerical format, enabling effective training of the neural network for burnout prediction.

Target Variable Encoding

The target variable, Burnout Symptoms, indicates whether an employee is experiencing burnout and is initially represented as categorical values (Yes/No). Since machine learning models require numerical input, this variable is converted into binary form, where Yes is encoded as 1 (burnout present) and No as 0 (no burnout). This transformation allows the neural network to process the target variable effectively during training and prediction.

The conversion can be represented as follows:

Burnout Symptoms	Encoded Value
Yes	1
No	0

It is upon this transformation that the machine learning model is able to view the target variable as a binary classification problem and as such the model learns to predict whether an employee is a member of the burnout category or not.

The following Python code may be used to implement the encoding process:

DF["BURNOUT SYMPTOMS"] =DF["BURNOUT SYMPTOMS"].MAP(YES:1, NO:0)

This command is used to find the numerical values of the original categorical values in other words, transforms the target column to a binary variable that can be used to train the model.

This encoding of the target variable will make sure that the neural network will be able to decode the output values in a proper way in the course of learning. It also makes the prediction easier so that the model has two categories of employees; burnout present (1) or burnout absent (0). The step is necessary to develop a successful machine learning model to predict the symptoms of employee burnout.

Feature Scaling

Feature scaling is an essential preprocessing step in machine learning and deep learning, as different features in a dataset often have varying numerical ranges. If these differences are large, models like Artificial Neural Networks may become biased toward features with higher magnitudes, negatively affecting learning and prediction accuracy. In the corporate stress dataset, variables exist on different scales, making scaling necessary to ensure all features contribute equally during training. This process improves model performance, stability, and convergence by standardizing the range of input values.

Feature	Example Range
Age	20 – 60
Monthly Salary (INR)	20,000 – 200,000
Working Hours Per Week	30 – 80
Sleep Hours	3 – 9

Some features such as Monthly Salary frequently resemble other such features, such as Sleep Hours or Stress Level, with values that are far smaller and so make the neural network assign them disproportionate weight during training. To avoid this disproportion, StandardScaler is used to scale the features, so that all of them are brought on the same scale. The scaler calculates the average and standard deviation of each feature with the help of the fit method and then uses the transform method to normalize the data. This process transforms all the features so that they have a mean value of zero and standard deviation of one so that the model is able to learn fairly using all the variables which enhances the overall performance of the training.

The dataset had vastly varying range features before using feature scaling. Some examples include:

Feature	Approximate Range
Age	20 – 60
Monthly Salary	20,000 – 200,000
Sleep Hours	3 – 9
Stress Level	1 – 10

Unless the variables are scaled in terms of their features, large numerical variables such as salary might overpower the learning process and the model would leave small but significant features like sleep hours behind. Through standardization, every feature is reduced to a standard scale with a mean of approximately 0 and standard deviation of approximately 1 in order to ensure that each feature makes equal contribution during the training process. This enables the neural network to gain useful relationships free of the

inclination with the high-magnitude variables. On the whole, scaling enhances the stability of the model, its optimization and accuracy of predictions.

C. Train Test Split

After preprocessing and feature scaling, the dataset is split into training and testing sets to ensure proper model evaluation. Typically, 80% of the data is used for training, allowing the model to learn patterns and relationships, while the remaining 20% is reserved for testing its performance on unseen data. This separation helps prevent overfitting, where the model memorizes training data instead of learning general patterns. Overall, this approach ensures that the model can generalize effectively and provide reliable predictions on new data.

The split may be mathematically expressed as follows:

Training Data = 0.8 x N

Testing Data = 0.2 x N

Where N is the total observations of the data.

As there are about 50,000 records of employees in the dataset, the distribution is:

Dataset	Number of Samples
Training Set	40,000
Testing Set	10,000

The method of splitting the data is through the traintestsplit() function of the Scikit-Learn library, which is normally used in preparing datasets in machine learning projects.

Xscaled, in this case, is the standardized independent variables or features and y is the dependent variable (Burnout Symptoms in this case). testsize = 0.2 indicates that one must test on 20 percent of the data. Also, the value of the parameter random_ state = 42 makes sure that the split of the data is kept and can be reproduced each time the model is run.

The train-test split will guarantee that the neural network is trained on a big section of the dataset whereas the rest of the data will serve as unknown samples to give a fair judgment. The method enables the researchers to determine the extent to which the model is generalizable to new data. The size of a training set also contributes to the fact that the neural network will learn nonlinear relationships between the influences at the workplace and employee burnout symptoms that will be more trustworthy in their approximations.

D. Neural Network Architecture

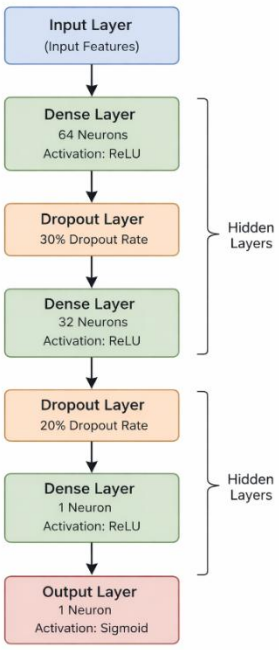


Fig 5: Architecture of Neural Network

Table 4.4.1: Dense Layer Configuration of the Proposed Model

Layer	Type	Units	Activation	Purpose
1	Dense	64	ReLU	Feature extraction
2	Dropout	—	—	Regularization
3	Dense	32	ReLU	Feature refinement
4	Dropout	—	—	Regularization
5	Dense	1	Sigmoid	Output prediction

A Feedforward Artificial Neural Network (ANN) is employed to forecast employee burnout by comprehending the connections between the variables at work and in personal lives. The model takes the form of an input layer that takes standardized features which include age, work pressure, job satisfaction, sleep duration and managerial support and then the two hidden layers with 64 and 32 neurons with ReLU that attempts to capture complex nonlinear patterns. The output layer has one neuron which has a sigmoid activation function, which gives a probability of burnout between 0 and 1. On the whole, this architecture allows the model to examine various factors at the same time and predict accurately whether burnout will occur or not based on learned patterns.

E. Mathematical Representation of the Neural Network

The workings of the applied neural network model involve each neuron receiving input features in the form of a weighted sum of a bias where the weights have been made to show the significance of the variables, work pressure, job satisfaction, commute time, sleep duration and managerial support, in predicting burnout. The model relies on ReLU activation in hidden layers in order to learn complex nonlinear relationships effectively since, unlike other superior models, it permits positive values to pass

through but filters out negative ones. A sigmoid activation function is used in the output layer to transform the final output into a probability of 0 to 1 to show the probability of burnout. The model predicts burnout or non-burnout based on a cut-off of 0.5, which allows a good binary prediction of workplace and lifestyle variables.

F. Neural Network Implementation

The neural network model was developed with the help of the high level of deep learning framework based on Python, TensorFlow Keras, which was used to build and train neural networks easily. The network architecture was built with a Sequential model in which every layer is added sequentially in a systematic way. The model uses this method to handle input features one after another in several layers until a final prediction is obtained.

Pseudo code:

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
model = Sequential()
model.add(Dense(64, activation='relu', input_dim=X_train.shape[1]))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

The model has an input layer and then a hidden layer with 64 neurons with activation of ReLU, which takes the input features and learns the preliminary patterns on the dataset. The second hidden layer of 32 neurons and ReLU activation is then included to identify more complicated relationships between the workplace factors and burnout symptoms. Lastly, a single neuron with Sigmoid activation is an output layer that produces a probability value that is the risk of the employee being burnt out.

Layers in the network are also realized by the Dense layers which are also known as fully connected Neural layers. During dense layer, every neuron is fed by all the neurons in the previous layer and therefore the model learns about the complex interactions between features. In the process of training, the network alters the values of the weight and the bias terms to reduce the prediction error, as well as, enhance its classification accuracy in respect to whether an employee suffers burnout symptoms given the input features

G. Model Compilation

The neural network is then compiled to make the training process known by defining the optimizer, loss function, and evaluation metrics before training. The model employs both Adam as an efficient and faster optimizer and Binary Cross Entropy as the loss function because the problem is a binary classification problem (burnout vs. non-burnout). The evaluation metric that is selected to evaluate correct predictions is accuracy. The loss function directs the minimization of the error during the training and the optimizer attempts to modify the weights, and the reduction of the loss value means that the model is learning and improving to predict more accurately.

H. Optimization Algorithm

The model of the neural net employs the Adam (Adaptive Moment Estimation) optimizer to estimate weights to update them effectively during the training process to achieve efficient stable learning. To obtain the benefits of both Momentum and the RMSProp, Adam modifies learning rate depending on the average and variance of the gradients, which enables faster and more stable convergence. This assists the model to modify parameters well during the process of backpropagation and find optimal solutions with less variations. Because of its rapidity, flexibility, and low requirement to tune the model manually, Adam will be suitable to train deep learning models to large and intricate datasets such as burnout prediction.

I. Model Training

Based on the standardized training dataset (Xtrain and ytrain), the neural network model gets trained and is able to acquire patterns and relationships between input features and the target variable, which is Burnout Symptoms. The model is trained with the help of the 35-epochs and 32-batches model.fit() function to guarantee an efficient and stable learning and validation is conducted with 20 percent of the data to be able to track the performance on the unseen data. At each iteration, the model produces predictions, computes error with the help of a loss function, and changes weights with the help of backpropagation and an optimizer to eliminate the error. This repetitive process allows the model to become more and more precise and also to succeed in capturing inherent regularities in the data.

```
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.9790 - loss: 0.0511 - val_accuracy: 0.9741 - val_loss: 0.0551
Epoch 12/35
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9791 - loss: 0.0489 - val_accuracy: 0.9764 - val_loss: 0.0559
Epoch 13/35
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9794 - loss: 0.0486 - val_accuracy: 0.9748 - val_loss: 0.0543
Epoch 14/35
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9795 - loss: 0.0488 - val_accuracy: 0.9750 - val_loss: 0.0554
Epoch 15/35
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9793 - loss: 0.0475 - val_accuracy: 0.9741 - val_loss: 0.0552
Epoch 16/35
328/328 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9809 - loss: 0.0465 - val_accuracy: 0.9746 - val_loss: 0.0544
352/352 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step
Final Accuracy: 97.60%
```

Fig 6: Model Training

The training log shows that your neural network is learning effectively and has converged to a high-performance state. As epochs progress (around epochs 12–16), both training accuracy (~97.9–98.0%) and validation accuracy (~97.4–97.6%) remain consistently high with very small fluctuations, indicating stable learning without overfitting. The loss values for both training (~0.046–0.051) and validation (~0.054–0.056) are low and closely aligned, further confirming good generalization. The minimal gap between training and validation metrics suggests that the model is not memorizing the data but capturing meaningful patterns. Early stopping likely prevented unnecessary training beyond this point, preserving optimal weights. Overall, achieving a final accuracy of 97.60% indicates that your model performs very well on this dataset, with strong predictive capability and good generalization to unseen data.

J. Results and Graph Analysis

Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	5619
1	0.98	0.97	0.98	5619
accuracy			0.98	11238
macro avg	0.98	0.98	0.98	11238
weighted avg	0.98	0.98	0.98	11238

Fig 7: Classification Report

According to the classification report, both classes (0 = non-burnout, 1 = burnout) are extremely well-performing in your model. In the case of class 0, the precision of 0.97 and recall of 0.98 indicate that the model is right most of the time in identifying the non-burnout-related cases without making many false positives. In the case of class 1, the accuracy of 0.98 and the recall of 0.97 mean that the performance is approximately the same with minimal false negatives to identify the burnout cases. F1-scores of the two classes are 0.98 showing a good balance of precision and recall. The equal support (5619 samples in each category) proves that the dataset is balanced and this factor also makes the model reliable in its performance. Also, 98 percent accuracy and the same macro and weighted averages (0.98) mean that the predictive ability of the model remains constant and predicts the classes accurately without discrimination. In general, the model has a high level of generalization and can be used to classify burnouts.

Burnout Class Distribution

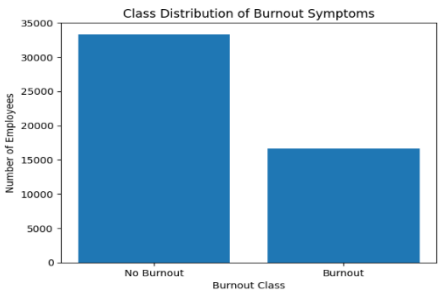


Fig 8: Class Distribution of Burnout Symptoms

It can be noted that the dataset has an evident bias in terms of the classes, with many more non-burnout employees than there are burnouts, which also increases exposing the majority patterns in the model in training. This disbalance could result in bias whereby the model can lean towards predicting non-burnout and fail to correctly predict employees at risk which can increase misclassification. Consequently, this implies that methods like SMOTE, class weighting or oversampling the minor population need to be used to equalize the sample and enhance the capability of the model to identify burnout cases accurately.

Confusion Matrix Analysis

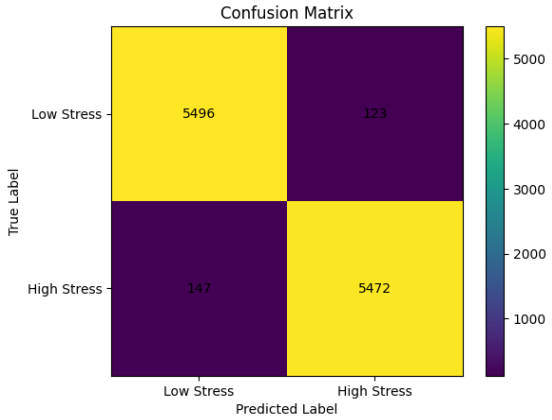


Fig 9: Confusion Matrix Distribution of Count Vs Components

The confusion shows that the model is very effective in distinguishing both low and high cases of stress, only 123 false positives and 5496 true negatives with 5472 true positives and 147 false negatives. The misclassifications are low, which means it is very accurate and a good predictor. The absence of minimal false positives will help prevent cases of employees being wrongly identified as burnt out, whereas low false negatives will help to identify the majority of high-risk persons. On the whole, such a sensitivity and specificity allow stating that the model is sufficiently reliable and can be applied to the actual situation to identify the risk of burnout.

K. Feature Correlation Analysis

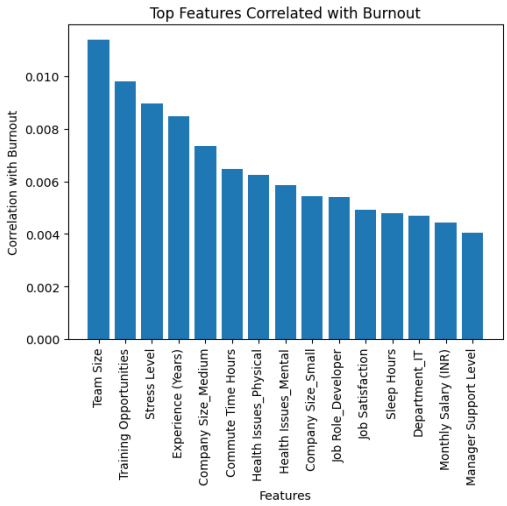


Fig 10: Top Features Correlated with Burnout

Correlation analysis determines the highest features that are most closely related to employee burnout that comprise the stress level, training opportunities, team size, years of experience, company size, commute time, physical or mental health issues. Of these, stress level has the greatest direct influence; as more stress a person is likely to burn out, whereas limited opportunities of training are related to lesser career development and job dissatisfaction. Commuting time is also a contributing factor to fatigue and inadequate work-life balance, which increases the risk of burnout, and employees with already existing health problems are more susceptible to burnout because they are unable to cope with pressure at work. In general, the findings are useful to prove that the issue of burnout is conditioned by a complex of

organizational and individual factors, and the ability to take into account a variety of variables is crucial in treating the issue of burnout predictability and in the management of stress at work.

L. Neural Network Feature Importance

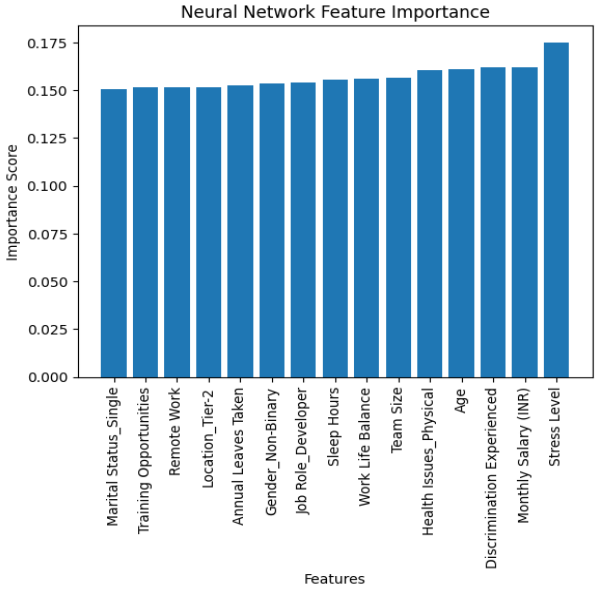
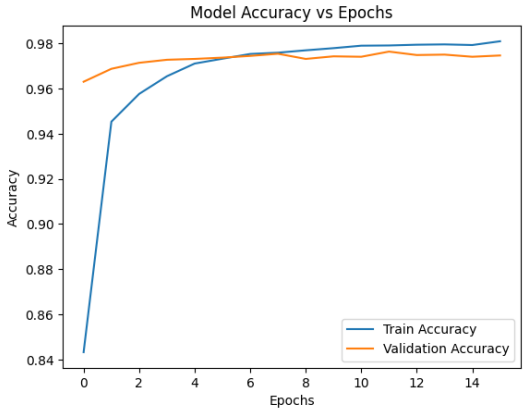


Fig 11: Neural Network Feature Importance for Scope Vs Features

The correlation analysis shows the strongest associated features with employee burnout such as the level of stress, training opportunities, the size of the team, the length of experience, the commuting time, the size of the company, the physical or mental condition. Among them, stress level demonstrates the most direct relationship, and higher levels of stress pose a great threat of burnout, whereas a lack of training opportunities and long commuting times lead to dissatisfaction, fatigue, and poor work-life balance. Also, the workers who have pre-existing health problems are also at higher risk of burnout since they find it harder to cope with the pressure at work. All in all, the findings prove that the factors that lead to burnout are a mixture of organizational and personal, and the multi-dimensional methods should be used in both prediction and management.

M. Epoch vs Loss Trend



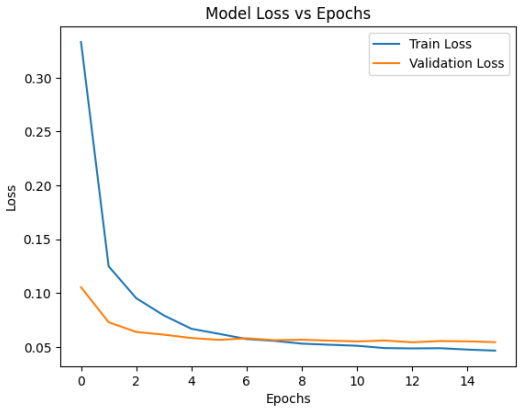


Fig 12: Epoch vs Loss Trend Analysis

Model performance is positive and steady as depicted in the accuracy and loss graphs in training. The model demonstrates fast learning during the early epochs whereby the accuracy increases dramatically between 84-94 percent to 97-98 percent both in the training and validation phase indicating successful learning and good generalization. In the same manner, the loss is greatly reduced by approximately 0.33 to 0.06 and is practically similar to both training and validation, which means that there is a low degree of overfitting and stable pattern learning. Comprehensively, the fact that the model converges smoothly and yields stable curves is an indication that the model is optimized and that it predicts the workplace factors and burnout symptoms in a reliable way.

N. Overall Model Observations

The results of the experiment also reveal that neither a single factor of workplace nor a personal factor affects employee burnout, but rather, a combination of the work and personal factors, such as stress levels, work-life balance, health, salary, and organizational environment. The model has high learning performance, and early epochs are high, followed by constant convergence as the training advances, and both the accuracy and the loss are high (approximately 97-98 percent) and low (approximately 0.05). The near correspondence between the training and validation outcomes indicates the presence of good generalization and a low degree of overfitting which proves the reliability of the model in the unseen data. All in all, the model demonstrates stable, consistent, and correct performance, which is why it is efficient in predicting burnout in employees.

O. Final Result Summary

The experimental results demonstrate the effectiveness and robustness of the proposed burnout prediction model. Using a dataset of 50,000 employee records with 29 features, an Artificial Neural Network (ANN) with two hidden layers (64 and 32 neurons) was implemented to capture complex patterns in the data. The model was trained over 35 epochs using backpropagation, enabling it to learn relationships between workplace factors, lifestyle habits, and burnout indicators. Evaluation on unseen test data showed a high accuracy of approximately 97–98%, along with low loss and stable validation performance, indicating strong reliability and good generalization capability of the model.

Table 4.15.1: Model Comparison

Model	Accuracy	Key Characteristics
Artificial Neural Network (Proposed Model)	97–98%	Captures complex non-linear relationships, high generalization
Logistic Regression	85–88%	Simple, interpretable but limited for complex patterns
Random Forest	90–93%	Good performance, handles non-linearity but less optimized than ANN

The model has a high level of practical utility in the organizational context since it allows to identify employees at risk of burnout as soon as possible to deliver proactive measures aimed at enhancing well-being, productivity, and efficiency at the workplace. Prediction functionality was introduced which preprocesses input data, scales it and yields an approximation which is a probability score which is categorized to High or Low Stress Risk on a 0.5 cut off point. According to test situations, it is observed that profiles with balanced working hours, adequate sleep and high satisfaction with the job are correctly classified as low risk whereas those having long working hours, poor sleep, high pressure and low satisfaction are classified as high risk. The results prove the effectiveness of the model in its generalization and indicate its potential in the field of real-life implementation in stress monitoring and early burnout prevention.

V. Conclusions

The purpose of the study was to come up with an evidence-based model that could be used to estimate stress levels among working professionals based on lifestyle, occupational, and demographic data and overcome the shortcomings of the traditional subjective data-gathering tools such as surveys. It found a void in scalable computational models and trained machine learning algorithms and approaches, such as Random Forest, Logistic Regression and Artificial Neural Network, on a large sample of 50,000 people to examine patterns of stress. Results have shown that stress is characterized by several issues that include the working hours, sleep, work-life balance, lifestyle habits, and organizational environment, and the ANN model made an approximate 97 percentage in determining stress levels. Although the model cannot encompass all the psychological and environmental variables, it could be useful in the early stress diagnosis and interventions in organizations. The paper identifies the promise of computational mental health based on data and recommends the research to be further enhanced by methods such as balancing data sets, using sophisticated feature engineering, and incorporating real-time behavior and physiological information.

Acknowledgment

The authors would like to express their sincere gratitude to all individuals and organizations who contributed to the successful completion of this research work. We extend our appreciation to our mentors

and academic guides for their continuous support, valuable suggestions, and insightful guidance throughout the study. Their expertise and encouragement played a crucial role in shaping this research.

We are also thankful to the developers and contributors of the TensorFlow and Keras frameworks for providing powerful tools that facilitated the implementation of the machine learning and deep learning models used in this work. Additionally, we acknowledge the NHANES data portal and the Centers for Disease Control and Prevention (CDC) for making high-quality and reliable datasets available for research purposes.

we would like to thank our peers, friends, and family members for their constant motivation and support, which helped us successfully complete this research paper.

REFERENCES

[1] Al-Alim, M. A., Mubarak, R., Salem, N. M. & Sadek, I., 2024. A machine-learning Approach for Stress Detection Using Wearable Sensors in Free-living Environments, s.l.: Helwan University.

[2] Alharbe, N. R., 2025. Soft computing analysis of the factors associated with stress, anxiety, and depression. BMC Public Health, pp. 1-13.

[3] Alruily, M., 2023. Sentiment analysis for predicting stress among workers and classification utilizing CNN: Unveiling the mechanism. Alexandria Engineering Journal, Volume 81, pp. 360-370.

[4] Breuer-Asher, I. et al., 2024. Association of Digital Engagement With Relaxation Tools and Stress Level Reduction: Retrospective Cohort Study. JMIR FORMATIVE RESEARCH, Volume 8, pp. 1-12.

[5] Chandrasekaran, S., Guduru, R. & Loganathan, S., 2025. Factors causing work related stress and strategies for stress management: a study of working women in private and public sectors in the Indian context. Frontiers in Global Women's Health, pp. 1-8.

[6] Dhanalakshmi & Dev, N., 2024. AN EVALUATION OF EMPLOYEES PERCEPTIONS OF WORKPLACE STRESSORS AND SOLUTIONS: DATA FROM THE SOFTWARE SECTOR. Journal of Lifescience and SDG's Review, pp. 1-13.

[7] Filippis, R. d. & Foysal, A. A., 2024. Comprehensive analysis of stress factors affecting students: a machine. Discover Artificial Intelligence, pp. 1-17.

[8] Haque, Y. et al., 2024. State of the Art of Stress Prediction from Heart Rate Variability Using Artificial Intelligence. Springer- Cognitive Computation, pp. 455-482.

[9] Hosseini, S. et al., 2022. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Scientific Data, Volume 25, p. 9.

[10] Jukic, T. et al., 2020. The effect of active occupational stress management on psychosocial and physiological wellbeing: a pilot study. BMC Medical Informatics and Decision Making, Volume 20, pp. 1-8.

[11] Lazarou, E. & Exarchos, T. P., 2024. Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices. Neuroscience, 11(2), pp. 76-102.

[12] Mahajan, A., Desai, I. P. & Muley, A., 2025. Assessing the Lifestyle related Determinants among Employees Working in the IT Sector of Pune City. Indian Journal of Community Medicine, 50(2), pp. 1-6.

[13] Masri, G. et al., 2024. Mental Stress Assessment in the Workplace: A Review. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, 15(3), pp. 1-19.

[14] Matsumoto, K., Matsui, T., Suwa, H. & Yasumoto, K., 2023. Stress Estimation Using Biometric and Activity Indicators to Improve QoL of the Elderly. Sensors, Volume 23, pp. 196-208.

[15] Mitra, P. & Sharma, P. (. R., 2025. A Descriptive Study of Work Life Balance, Perceived Stress and Coping Styles in Male and Female Educators. The International Journal of Indian Psychology, 13(1), pp. 1-17.

[16] Pabreja, K. et al., 2021. Prediction of Stress Level on Indian Working Professionals Using Machine Learning. International Journal of Human Capital and Information Technology Professionals, 13(1), pp. 1-26.

[17] Patil, V. C., Patil, S. V., Shah, J. N. & Iyer, S. S., 2021. Stress Level and Its Determinants among Staff (Doctors and Nurses) Working in the Critical Care Unit. Indian Journal of Critical Care Medicine, 25(8), pp. 887-891.

[18] Shahapur, S. S. et al., 2024. Decoding Minds: Estimation of Stress Level in Students using Machine Learning. INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY, pp. 2002-2013.

[19] Trivedi, O. et al., 2024. Levels of work stress among information technology professionals during COVID 19 pandemic in an Indian metropolis. Journal of Family Medicine and Primary Care, pp. 674-681.