Frontiers in Health Informatics www.healthinformaticsjournal.com
ISSN-Online: 2676-7104

2023; Vol 12 Open Access

DECODING STRESS BEHAVIOR: A CLUSTERING-BASED ANALYSIS OF WORK–LIFE BALANCE AND DIGITAL HABITS

Zahid Hussain Wani[1,*], Hilal Ahmad Khanday[2],
[1,2] Department of Computer Science, University of Kashmir, India.

Corresponding Author: *zahid.uok@gmail.com***Abstract** - *The major source of significant stress is the association between digital habits, work-life management, and mental strain. Although major research relies on direct prediction, this approach could simplify the complicated as well as varying nature of stress. This research uses a large dataset (N = 24,045) having 18 behavioral, demographic, and physiological aspects and applies an unsupervised learning algorithm to analyze it. A K-means cluster algorithm was used to identify unique user profiles without any presumptions of their stress level, then tested using a random forest classifier for statistical robustness, and characterized via descriptive statistics. The investigation successfully recognized three separate, strongly distinguishable behavioral archetypes: (1) balanced professional life, the low stress, identified by high job satisfaction, and good sleeping routines; (2) the overworked and stressed, characterized by frequent exhaustion, high stress, and extended work time; and (3) the digitally fully engaged, distinguished by notification overload along with excessive social media. The classification accuracy of 93% ensured the integrity of the cluster. The study suggests that for better stress-related insights and personalized wellness, a person-centered profile is a better approach than typical assumption models.*

Keywords— Behavioral Archetypes; Digital Well-being; Cluster Analysis; Machine Learning, Unsupervised learning; Work-Life Balance

Introduction

Stress has grown into an unavoidable aspect of the daily grind. Ultimately, it is recognized as a striking issue in the field of health [1]. Prolonged stress leads to negative influence on high blood pressure, sleep difficulties, increased risk of heart attacks, and so on [2]. Stress can be caused by any aspect of either external or internal factors. Demands as well as pressure at the workplace result in external stress, and imbalanced emotions lead to internal stress. Furthermore, fear, future concerns, self-pity, insufficient scheduling, intense job expectations, and workplace insecurity are prevalent issues and sources of stress. Additionally, the contribution of health problems and sleep abnormalities is also in negative stress [3]. Stress contributes significantly to psychological as well as physical disorders in individuals [4]. According to the American Institute of Stress, the second half of employees require assistance in handling stress, while 45% think others around them also require it. 80% of staff members experience distress in their positions. 54% of all employees took off from work, and 44% of all cases in which persons endured health-related issues were caused by stress, depressive disorders, and anxiousness in 2018 as well as 2019, as per the Health and Safety Executive (HSE) [5]. Reports by the Chinese Center for Disease Control and Prevention show that suicide is now the primary cause of mortality among adolescents, with chronic stress being the significant factor. In compliance with all points, a swift increase in stress is a threat to human well-being along with lifestyle [6].

Multiple approaches have been proposed for measuring occupational stress. Self-reported stress, also known as perceived stress, is frequently used as a baseline for assessment on scales such as the Perceived Stress Scale, Daily Stress Inventory, and single-item surveys. However, these approaches are prone to memory and affect biases and therefore are ineffective for continuous monitoring. With advancements in wearable sensors, scientists are now turning to unobtrusive, real-time stress monitoring using physiological signals, particularly when computing [7]. Recently, various models have been trained on the digital data set utilizing machine learning for stress forecasting [8] like other forecasting problems. For instance, distress prediction associated with stress has been detected using several mobile sensory data [9,10]. Furthermore, by examining post patterns on social media platforms, stress has been predicted [11].

While machine learning techniques frequently neglect clear demographic background as well as explicit user-reported data on habits. With the creation of methodology built around a hybrid dataset incorporating user demographics, platform choices, and proactive healthcare behaviors, the gap is closed by this study. Identification of behavior's impact on stress is difficult since it may arise from various situations. This research paper focuses on using an unsupervised cluster approach to recognize several behavioral patterns, often referred to as behavioral archetypes, in individuals. An examination of a broad collection of data, including work habits, social media use, and health routines, revealed three unique user types. This study shows that, rather than attempting to forecast stress directly, classifying people based on their behavior might offer us a more comprehensive picture of how stress presents itself in everyday life.

The research is directed by two main objectives:

- Using unsupervised machine learning on a real data set, determine different behavioral archetypes determined by work habits, digital media use, and wellness routines.
- For an understanding of what separates each archetype, measure essential attributes such as work hours, sleep, and job satisfaction.

literture review

In recent years, the increased prevalence of psychological stress associated with fast-paced, technology-driven lives has emerged as a major public health problem. As a result, academics are increasingly using machine learning and artificial intelligence (AI) to better identify, predict, and manage stress. A deep learning-based Mental Health AI, utilizing various factors in order to predict stress, was developed by [12]. A fully connected neural network (FCNN) operates in the VADER NLP framework to get information about demographics, pulse rate data, feedback from users, and online platforms provided sentiment evaluations. It helped a lot in evaluating the mental health to the depth as well as the method involved advanced intervention approaches. Study in [13] trained a deep learning model on 270 recorded calls to measure stress via emotional trends within voice-based connections. The recordings were split into 27,000 audio samples, and the model was forecasting stress 80% correctly. Afterwards it was improved through information gathered from 40 staff as well as 4,672 customer conversations. [14] designed an interface to anticipate future stress levels via data from wearable devices such as heart rate. The random forest model was applied to train a model on a dataset containing 400 entries and afterwards tested on one of data gathered from two employees. Therefore, results show an effective correlation, that is, $r = 0.793$ and 90% of the time error was almost less than 12 BPM. Recent [15] has demonstrated a stress-predicting model using artificial intelligence (AI) along with machine learning (ML) models, with the help of historical trends based on job as well as personal parameters. It works on data of employ along with providing active monitoring. Dham et al. [2] used different models to predict stress based on available data related to physiology; as a result, the stacking classifier with 99.92% outperforms all other models. As a result, citing works by Sevil et al. [16] and Ding et al. [17] was exceptionally reliable along with major improvements. Jaber *et al.* [18] created AI system based on a physiological data set, further applied random forest classifier, achieved 0.78 as F1-score, which is comparable with other models, for stress foreseeing. Tzirakis et al. [19] used a neural network model to predict stress via processed video footage. Mirsamadi et al. [20] detected stress by analyzing speech using recurrent neural network.

In conclusion, the current literature provides a rapidly growing foundation for identifying stress through advanced machine learning as well as artificial intelligence methods using various sources like wearables, voice recordings, and social media. For all studies distinct patterns are used, which represents majorly two main characteristics:

1. *A focus on direct prediction:* The overarching goal of the examined research, which includes Ramu et al. (2023) to Jaber et al. (2021), aims to establish high-accuracy model for prediction. Mostly, parameters like accuracy, RMSLE, and F1-score are utilized to measure the performance of the model.
2. *A Reliance on Biometric and Sensor Data:* The most common sources of data are signals from the body (HRV, ECG, GSR), videos, speech patterns, and verbal sounds analysis.

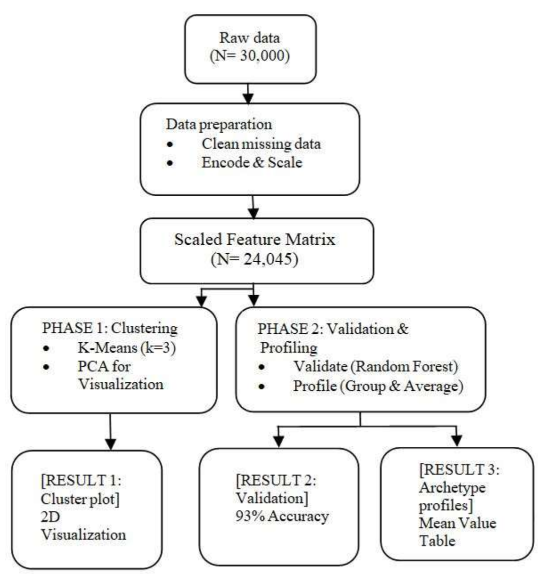
This is the particular need our study seeks to fill. Rather that estimating a stress score based on physiological signs, this work takes a person-focused, unsupervised learning method. Our objective in studying a separate collection of self-reported behavioral, vocational, and wellness variables is not to attain prediction accuracy in the usual sense, but rather to find and define the population's latent behavioral archetypes. This method aims to give a contextual understanding of stress, complementing the present literature's largely physiological focus, by moving the emphasis from immediate identification to holistic assessment.

methodology

The qualitative as well as sectional approach is adopted to investigate undiscovered patterns in human activity data along with its connection with stress. The evaluation is done through two phases, at first detection of different kinds of behaviour utilising clustering technique is done under unsupervised learning. Second is the explanation and validation phase, which includes studying these behaviour types elaborately, then analyse their reliability. The method described here more accurately represents the complexity in addition to the varied character of responses to stress without oversimplifying them into fixed categories [21].

Data source, feature description, and preprocessing

1) *Dataset:* Kaggle, an online platform that provides data on benchmarks [22], is utilized in this research for dataset collection. The selection was done on the basis of its rich parameter listed as important to our present workplace harmony as well as digital wellness.
2) *Feature space:* The collection includes 30,000 actual event records reflecting the personality traits of people across a variety of employment, social routines, and decisions regarding their lifestyles. Out of all 19 features for classification one character represents the result factor utilized named “stress_level”. Other 18 features are input variables,



in order to offer an in-depth understanding, they have been thematically put together into four distinct categories:

a) Demographic domain: “age” and “gender” are the basic factors.

Figure 1- Flowchart: Methodological Flowchart of the Study

b) *Occupational domain:* All work-related features like “job_type”, “work_hours_per_day”, “breaks_during_work”, “perceived_productivity_score”, “actual_productivity_score”, and “job_satisfaction_score”.

c) *Digital behavior domain:* Parameters that capture internet usage such as “daily_social_media_time”, “number_of_notification”, “social_platform_preference”, “screen_time_before_sleep”, and the consumption of digital wellbeing services (“uses_focus_apps”, “has_digital_wellbeing_enabled”).

d) Lifestyle and wellness domain: Individual health determinants include “sleep_hour”, coffee_consumption_per_day”, “weekly_offline_hours”, and “days_feeling_burnout_per_month”.

3) Data preprocessing: The data’s consistency was maintained by eliminating each entry containing one or more missing values (NaN) by using the listwise deletion method [23]. Even though sample size is reduced, but consistency as well as completeness is guaranteed for future statistical models [24], alongside minimizing the biasness caused by data interpolation. Following preprocessing, offering a solid platform for statistical analysis, the sample set totaled 24,045 individuals.

Data transformation and preparation for modeling

The raw data was transformed to make it suitable prior to implementation of the machine learning model, as only numerical and standardized inputs are valid [25].

1) Encoding of categorial variables: The data was converted from qualitative format into machine-understandable data. Scikit-learn’s LabelEncoder was used for transforming features “gender”, “job_type”, “social_platform_preference” into unique integers for each category within features. Furthermore, boolean features (“uses_focus_apps”, “has_digital_wellbeing_enabled”) were converted to 0 (false) or 1 format (true).

2) Feature standardization: Due to the usage of the distance-based technique K-means, it is necessary to arrange data by feature scaling [26]. By using StandardScaler from scikit-learn, all 18 columns were standardized. The process involves conversion of each feature by subtracting their mean, then dividing by their standard deviation, producing 0 and 1 as mean and variance, respectively [27]. The features with naturally greater scales (for example, “weekly_offline_hours”) are required to be standardized, otherwise it may result in having an excessive impact when contrasted with smaller-scale features while clustering.

Phase 1: unsupervised discovery of behavioral archetypes

The primary goal of phase 1 was to split the test group into unique as well as internally cohesive sections using unsupervised learning, despite any previous understanding of their stress level.

1) Clustering algorithm rationale: As per the primary algorithm, the K-Means clustering method was chosen. K-Mean is iterative in nature and a centroid-based technique recognized by its efficiency alongside scalability on huge datasets [28][29]. It was aimed to partition n rows into k clusters, where every observation was assigned to the cluster with the closest mean, which is considered as cluster’s prototype [30]. The reduction of the within-cluster sum of square (WCSS) was our main objective.

2) *Cluster number specification:* The n has value 17074, and k (number of clusters) was specified as 3 to summarize general patterns. Furthermore, it also ensures that identified behavioral archetypes are distinct enough while being meaningful as well as over-segmentation can be avoided.

3) *Visualization via dimensionality reduction:* Principal Component Analysis (PCA) was implemented in order to explain and illustrate highly dimensional clustering conclusions. It is a linear decrease in dimensionality approach that converts data into a fresh coordinate system while the principal components capture the greatest variation [31]. A 2D scatter plot was applied to examine cluster divisions after limiting the 18 features to two important components. PCA was only used for visualization, but the K-Means algorithm was trained on the whole 18-dimensional standardized features, so that all information was extracted for segmentation.

Phase 2: archetype validation and profiling

After identifying the three clusters, a phase was dedicated to establishing their statistical validity and developing descriptive profiles to better comprehend their significance.

1) *Quantitative cluster validation:* The topic was presented as a classification with multiple classes of jobs in order to validate the cluster. Considering high accuracy and resistance to overfitting, the random forest classifier fits best [32] to predict an individual's cluster grouping using 18 behavioural as well as demographic parameters. The assumption behind this validation is that this supervised model ought to be capable of predicting membership with outstanding precision only if clusters are actually different as well as clearly distinct in the field of features [33] . As a result, classification performance could potentially be used to estimate the statistical stability of the clustering approach.

2) *Qualitative cluster profiling:* A descriptive statistical study was conducted to convert the numerical groupings into significant behavioural archetypes. Based on their given label and the mean score for each of the 18 characters as well as the "stress_level" clusters, were formed. As a result, a thorough profile for each archetype is generated, allowing for comparison study so that distinct blend of high, medium, and low feature scores that distinguish and separate each group from one another can be found.

Technical implementation

Python was leveraged for data analyzing, modeling purposes, and visualization. Various open-source scientific computation libraries such as pandas (for data management and manipulation), scikit-learn (for preprocessing and machine learning models like K-Means, PCA, and Random Forest Classifier), and Matplotlib along with seaborn (for excellent data visualization) were implemented.

result and discussion

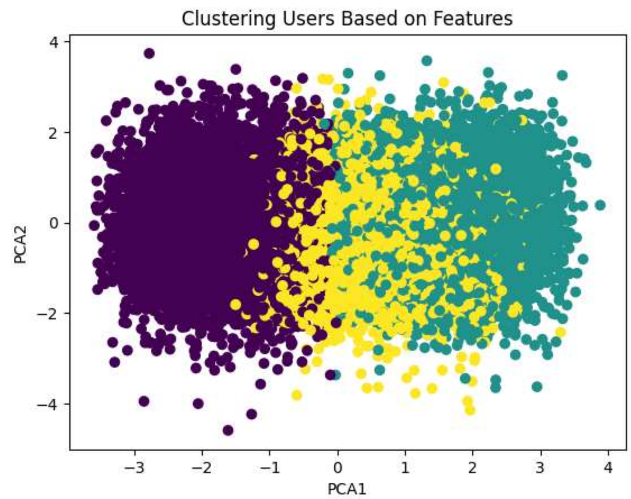
The main findings from a sample of 24,045 people are described in two parts: the use of unsupervised learning for behavioral archetype identification and the validation along with a comprehensive definition of these archetypes.

A. PHASE 1: Identification of behavioral archetypes via clustering

The primary study involves implementing the K-Means technique to all the standardized 18-feature data, so that naturally existing categories of users within the sample data can be identified.

1) *Cluster detection:* The algorithm successfully divided the final sample of 24,045 individuals into three unique groups, as shown in figure. The distribution of persons among these clusters has been determined to be:

a. Cluster 0: 8,416 individuals (35.0%)
b. Cluster 1: 9,185 individuals (38.2%)
c. Cluster 2: 6,444 individuals (26.8%)



2) Visualization-based cluster distinction: The 18-dimensional feature spaces were subjected to Principal Component Analysis (PCA) in order to show the separation of these groups. Figure shows three clusters shown across the first two main components. The figure shows a significant visible demarcation between the clusters, notably the huge, dense Cluster 1, and the remaining two clusters, giving strong visual proof that the algorithm detected distinct basic trends in the data.

B. PHASE 2: Archetype Validation and Characterization

Figure: PCA Plot of the Three Behavioral Archetypes

A validation alongside profiling evaluation was conducted to check the statistical integrity of clusters and to further understand their identity properties.

Quantitative Cluster Validation: *Using Random Forest classification, all three clusters were evaluated using just their 18 input characteristics, and validation test accuracy came out to be 93%. Indicating that the clusters are statistically resilient and highly separable groupings in the data.*

Archetype Characterization: *After validation, each archetype was defined on the basis of its average attributes. Table aggregates the mean readings for all characteristics across the three groups and reveals three unique behavioral archetypes.*

Table: Mean Feature Values for Each Identified Archetype

Feature	Archetype 0: The Balanced Professional	Archetype 1: the Overworked & Stressed	Archetype 2: The Digitally Immersed
stress_level	2.98	7.45	5.51
work_hours_per_day	7.05	9.52	7.63
job_satisfaction_score	7.51	3.03	4.88
sleep_hours	7.79	5.55	6.49
weekly_offline_hours	24.89	9.98	12.15
days_feeling_burnout_per_month	3.88	19.95	10.02
daily_social_media_time	1.85	2.11	5.05
number_of_notifications	39.8	45.1	71.3
screen_time_before_sleep	0.75	1.15	2.01
coffee_consumption_per_day	1.49	3.48	2.55
breaks_during_work	7.1	2.5	5.5
perceived_productivity_score	7.45	3.99	5.11
actual_productivity_score	7.51	4.05	5.03

uses_focus_apps	0.65	0.21	0.15
has_digital_wellbeing_enabled	0.71	0.15	0.25
age	40.1	39.5	39.8
gender	0.51	0.49	0.50
job_type	3.55	3.48	3.51
social_platform_preference	2.99	3.01	2.95

Table gives a clear narrative for three unique behavioral archetypes identified in this study by providing analysis of the mean feature values for each cluster. The results show that anxiety is not a single experience but rather is profoundly rooted in overall lifestyle patterns.

a. Archetype 0: The Balanced Professional
The lowest stress level (2.98) is represented by this section, reflecting the highest degree of well-being. This excellent outcome is backed by a set of healthy habits: they have the greatest score for job satisfaction (7.51), the longest duration of sleep (7.79 hours), and the most breaks at work (7.1). Furthermore, they have maximum weekly online hours that are 24.89, which implies they lived a life of substantial detachment. Also resulted in the lowest rate of burnout per month. This typology is not passively healthy but actively involved in its own well-being, as seen by their proclivity to employ focus applications and digital well-being products.

b. Archetype 1: The Overworked and Stressed
In sharp contrast, this category has the highest stress level (7.45), representing a high-risk character for occupational burnout. The data shows that their work life is the key driver: they work the most hours (9.52 per day), experience the least satisfaction with their jobs (3.03), have the shortest breaks (2.5), a troubling 19.95 days of burnout each month, and have shortest amount of sleep (5.55 hours). Notably, their social media time is not excessive, but coffee intake is excessive, which means their lifestyle is driven by the need to compensate for tiredness. Therefore, stress is primarily work-centric but not digitally driven.

c. Archetype 2: The Digitally Immersed
This archetype is of moderate stress level (5.51) which is in between the other two types. They have a relatively higher level of digital engagement, with an average of 71 notifications received per day during the past week, compared to other archetypes, and spending 5+ hours on social media per day. This "always-on" digital lifestyle extends into the night, with more than two hours of screen time before retiring to sleep, that may be leading to their average sleep and stress scores. The indicators they have to work are mild, meaning that this group's salient distinction is that they have absorbed modern technology to such an extent that they do not necessarily have to work hard to do so.

It may be concluded that there are several factors with stress and not just a single attribute. The data clearly separates stress caused by an unsustainable work life (archetype 1) from constraint linked with high digital engagement (Archetype 2), alongside offering a clear blueprint for good well-being (Archetype 0). This allows for the shift from general recommendations to personalized wellness treatments based on a person's profile.

conclusion

To sum up, this research shifts the focus of analysis from prediction to full characterization, by reformulating our understanding of contemporary psychological stress. Using unsupervised machine learning we found three statistically distinct and very different behavioral archetypes: Balanced Professional, Overworked and Stressed, and Digitally Immersed, indicating the dependence of the human life on different factors and not on a single one. This method helps not only to determine if there are risk factors present, but it also captures the essence of different lifestyles. Therefore, resulting in a more profound and effective manner to assist individuals. Acknowledging the difference between related to work stress and digital exposure based stress makes a realistic approach to mental health therapy. It supports people, businesses and healthcare providers to develop efficient solutions which are tailored to the lifestyle of persons in the modern technologically advanced world.

References

[1] Seaward, B. L. (2019). *Essentials of managing stress*. Jones & Bartlett Learning.

[2] V. Dham, K. Rai, and U. Soni, "Mental stress detection using artificial intelligence models," in *J. Phys.: Conf. Ser.*, vol. 1950, no. 1, p. 012047, IOP Publishing, 2021.

[3] S. Gedam and S. Paul, "A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques," in *IEEE Access*, vol. 9, pp. 84045-84066, 2021, doi: 10.1109/ACCESS.2021.3085502.

[4] G. P. Chrousos, "Stress and disorders of the stress system," *Nat. Rev. Endocrinol.*, vol. 5, no. 7, pp. 374–381, Jul. 2009, doi: 10.1038/nrendo.2009.106.

[5] P. Bobade and M. Vani, "Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2020, pp. 51-57, doi: 10.1109/ICIRCA48905.2020.9183244.

[6] S. Hashemi, F. Ghazanfari, F. Ebrahimzadeh, S. Ghavi, and A. Badrizadeh, "Investigate the relationship between cell-phone over-use scale with depression, anxiety and stress among university students," *BMC Psychiatry*, vol. 22, no. 1, p. 755, Dec. 2022, doi: 10.1186/s12888-022-04419-8.

[7] F. Akbar, G. Mark, I. Pavlidis, and R. Gutierrez-Osuna, ‘ An Empirical Study Comparing Unobtrusive Physiological Sensors for Stress Detection in Computer Work ’, *Sensors*, vol. 19, no. 17, Art. no. 17, Jan. 2019, doi: 10.3390/s19173766.

[8] Y. Fukazawaa, T. Ito, T. Okimura, Y. Yamashita, T. Maeda, and J. Ota, "Predicting anxiety state using smartphone-based passive sensing," *J. Biomed. Inform.*, vol. 93, May 2019, Art. no. 103151.

[9] Y. Acikmese and S. E. Alptekin, "Prediction of stress levels with LSTM and passive mobile sensors," *Procedia Comput. Sci.*, vol. 159, pp. 658–667, 2019, doi: 10.1016/j.procs.2019.09.221.

[10] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Inform.*, vol. 92, p. 103139, Apr. 2019, doi: 10.1016/j.jbi.2019.103139.

[11] S. Chandra Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, and L. H. Ungar, "Understanding and measuring psychological stress using social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, no. 1, pp. 214–225, Jul. 2019.

[12] Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translational psychiatry*, *10*(1), 116.

[13] S. Bromuri, A. P. Henkel, D. Iren, and V. Urovi, "Using AI to predict service agent stress from emotion patterns in service interactions," *J. Serv. Manag.*, vol. 32, no. 4, pp. 581–611, Jun. 2021.

[14] V. A. Stefanescu and I. E. Radoi, "Stress level prediction using data from wearables," in *Proc. 2019 18th RoEduNet Conf.: Networking in Education and Research (RoEduNet)*, Oct. 2019, pp. 1–6.

[15] Abell, M. (2016). *Traditional EEG measures and temporal patterns of neural network synchronization in veterans with and without mild traumatic brain injury and posttraumatic stress disorder* (Doctoral dissertation, Indiana University).

[16] M. Sevil, M. Rashid, I. Hajizadeh, M. R. Askari, N. Hobbs, R. Brandt, M. Park, L. Quinn, and A. Cinar, "Discrimination of simultaneous psychological and physical stressors using wristband biosignals," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105898, Feb. 2021.

[17] Y. Ding, Y. Cao, V. G. Duffy, Y. Wang, and X. Zhang, "Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning," *Ergonomics*, vol. 63, no. 7, pp. 896–908, Jul. 2020.

[18] D. Jaber, H. Hajj, F. Maalouf, and W. El-Hajj, "Medically-oriented design for explainable AI for stress prediction from physiological measurements," BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 38, Feb. 2022.

[19] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Oct. 2017.

[20] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[21] T. Iqbal, A. Elahi, W. Wijns, and A. Shahzad, "Exploring unsupervised machine learning classification methods for physiological stress detection," *Front. Med. Technol.*, vol. 4, p. 782756, Mar. 2022.

[22] Banachewicz, K., & Massaron, L. (2022). *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd.

[23] Livera, A., Theristis, M., Koumpli, E., Theocharides, S., Makrides, G., Sutterlueti, J., ... & Georghiou, G. E. (2021). Data processing and quality verification for improved photovoltaic performance and reliability analytics. *Progress in photovoltaics: research and applications*, *29*(2), 143-158.

[24] Molenberghs, G., Kenward, M. G., Aerts, M., Verbeke, G., Tsiatis, A. A., Davidian, M., & Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, *23*(1), 11-41.

[25] K. Modarresi and A. Munir, "Standardization of featureless variables for machine learning models using natural language processing," in *Computational Science – ICCS 2018*, Y. Shi *et al.*, Eds., Lecture Notes in Computer Science, vol. 10861. Cham, Switzerland: Springer, 2018, pp. 230–243. doi: 10.1007/978-3-319-93701-4_18.

[26] Raeisi, M., & Sesay, A. B. (2022). A distance metric for uneven clusters of unsupervised K-means clustering algorithm. *IEEE Access*, *10*, 86286-86297.

[27] Lee, D. K. (2020). Data transformation: a focus on the interpretation. *Korean journal of anesthesiology*, *73*(6), 503-508.

[28] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982. doi: 10.1109/TIT.1982.1056489.

[29] M. Capó, A. Pérez, and J. A. Lozano, "An efficient K-means algorithm for massive data," in *Proc. Int. Conf. on Artificial Intelligence and Soft Computing (ICAISC)*, Springer, 2018, pp. 271–281. doi: 10.1007/978-3-319-91253-0_25.

[30] Nie, F., Wang, C. L., & Li, X. (2019, July). K-multiple-means: A multiple-means clustering method with specified k clusters. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 959-967).

[31] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, Apr. 2016. doi:10.1098/rsta.2015.0202.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. doi: 10.1023/A:1010933404324.

[33] Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., & Müller, K. R. (2022). From clustering to cluster explanations via neural networks. *IEEE transactions on neural networks and learning systems*, *35*(2), 1926-1940.