Burnout Prediction and Analysis in Shift Workers: Counterfactual Explanation Approach

Ziang Tang[1], Zachary King[1], Alicia Choto Segovia[1], Han Yu[1], Gia Braddock[1], Asami Ito[2], Ryota Sakamoto[2], Motomu Shimaoka[2], Akane Sano[1]

***Abstract*—Shift work disrupts sleep and causes chronic stress, resulting in burnout syndrome characterized by emotional exhaustion, depersonalization, and decreased personal accomplishment. Continuous biometric data collected through wearable devices contributes to mental health research. However, direct prediction of burnout risk is still limited, and interpreting machine learning (ML) models in healthcare poses challenges. In this paper, we develop machine learning models that utilize wearable and survey data, including rhythm features, to predict burnout risk among shift workers. Additionally, we employ the DiCE (Diverse Counterfactual Explanations) framework to generate interpretable explanations for the ML model, aiding in the management of burnout risks. Our experiments on the AMED dataset show that incorporating rhythm features significantly enhances the predictive performance of our models. Specifically, sleep and heart rate features have emerged as significant indicators for accurately predicting burnout risk.**

***Index Terms*—burnout syndrome, counterfactual explanation, machine learning, risk prediction, shift workers, wearable devices**

I. Introduction

Shift work increases sleep disturbance and health problems including insomnia, decreased alertness, depression, fatigue, and chronic stress, leading to burnout syndrome [1]. Burnout has been defined as a "syndrome conceptualized as resulting from chronic workplace stress that has not been successfully managed" [2].

Previous Studies showed correlations between burnout and variables such as age, work conditions, and social support [3]. Studies identify burnout subtypes [4] and predict burnout using survey data, clustering, and neural networks [5]. Job stress and burnout contribute to disengagement and withdrawal. Circadian rhythm and job demand impact burnout and job dissatisfaction [6]. Disrupted circadian rhythms, caused by factors like night shift work, can negatively affect mental health, mood, and cognitive functions [7] [8].

Wearable devices provide abundant information collected continuously without disrupting daily activities and have been used for well-being and mental health studies. Some studies combine wearable sensors with other data to predict mood and well-being [9], [10], [11], [12]. While stress prediction is common, direct prediction of burnout risk is understudied.

Another issue prominent in predicting health and mental health problems is that the decisions are non-interpretable to health providers or the patient receiving care. One common approach to explain an ML model involves employing a simpler surrogate model that provides interpretable information, such as scores indicating the importance of various features [13]. However, they have a fundamental trade off between fidelity and interpretability. Typically, a highly faithful explanation tends to be complex and challenging to comprehend, whereas an interpretable explanation is often inconsistent with the model it intends to explain. Counterfactual explanations offer a solution to this paradox by identifying necessary modifications of input that would alter the model's decision [14].

In this paper, we (1) develop machine learning models using wearable and survey data including rhythm features to predict burnout risk in shift workers, and (2) generate counterfactual explanations to change high burnout risk to low risk and analyze the characteristics of the explanations that might help manage burnout risks.

II. Methods

A. Dataset

The AMED dataset consists of longitudinal data collected from 75 shift workers. (physicians N=6 and nurses N=69) working at the intensive care unit (ICU) and emergency room at Mie University Hospital in Japan. The objective of the study was to evaluate the effect of a 4-week physician-assisted internet-delivered cognitive behavioral therapy (CBT) program for shift workers. Participants were eligible if they were on a 3-shift schedule and the Pittsburg sleep quality index (PSQI $\geq$ 5) indicated poor sleep. Exclusion criteria include (1) a diagnosis of sleep disorders; (2) a diagnosis of psychotic diseases; (3) use of sleep medications; (4) pregnancy; (5) a history of contact dermatitis or other skin diseases with a high risk for skin disorders. The baseline period was 1 week, followed by an intervention period of 4 weeks [15]. During the study, participants wore a Fitbit Charge 3 that recorded steps, heart rate, and sleep continuously for 24 hours every day for the duration of the study. Fitbit data had missing values in heart rate (14.8%), steps (14.5%), and sleeping data (0.4%). We addressed this using linear interpolation, and zero and template imputation. During the pre-intervention period, participants were asked to refrain from accessing their activity and sleep data. Participants also received personalized sleep advice from physicians, once during week 1 of intervention, and 3 times a week during the last 3 weeks of interventions [15]. The self-report measures obtained during the study include a daily morning and evening survey that measured contextual data

This research was funded by the Japan Agency for Medical Research and Development (No. 19217687) and National Science Foundation (#18040167).

[1] Rice University, Houston, TX, 77005, USA. zt23,akane.sano@rice.edu

[2] Mie University, Tsu, Mie, 514-0102, Japan.

(caffeine and alcohol intakes, number of naps, overwork time, and work shift types).

To obtain the ground truth of burnout risks, the participants completed the Japanese Burnout Scale (JBS). The JBS questionnaire consists of 17 questions reported using a 5-choice Likert scale. From the 17 questions, three subscales of burnout syndrome are derived: emotional exhaustion (EE), depersonalization (D), and personal accomplishment (PA). Each of these subscales is scored by averaging the scores of their respective questions. From the 3 subscales, we categorized them as high (n=36) or low (n=5) risk based on the cutoff for quitting a hospital due to burnout (EE > 3.27, D > 2.06, PA < 2.49) [16]. If a participant scored high risk in any of the subscales, then we labeled the participant as high risk or low risk otherwise.

B. Daily Features

We computed 20 daily features from survey and Fitbit data, including [heart rate] maximum, minimum, mean, standard deviation, and sample entropy (measurement of the self-similarity of a sequence) of heart rate, resting heart rate, [sleep] onset, duration, efficiency, regularity [17], [steps] daily steps, entropy of inactive moments, Entropy of active moments, [surveys] number of naps, number of cups of caffeine drinks and alcoholic drinks, wake up type, shift types, work hours, and overwork hours. See [15] for feature engineering.

C. Rhythm Features

Biological processes often display oscillations that follow different biological clocks. Depending on the period of the oscillations, we can classify them into ultradian (less than 24 hours), circadian (24 hours), or infradian rhythms (greater than 24 hours). We first computed hourly features (hourly steps, heart rate mean, heart rate standard deviation, sleeping minutes) and then used this data to extract rhythm features such as (a) Most-active 10 h (M10) and least-active 5 h (L5), (b) deviation from a template (computed a 24-hour template by averaging the activity of each hour for each time window (7, 14, 21, and 28 days) days, and for each day, computed hourly differences between the day and the template. From the resulting value, calculated the mean, median, standard deviation, skewness, and kurtosis), (c) mesor, acrophase, period, good of fit from cosinor method using CosinorPy [18] for 8, 16, 20, 24, 28, 32, 36, 64, 72, 128, 256, and 512 h. See [19] for details of rhythm features.

The hourly granularity leaves room for many missing values. We first removed days where there is no data available for any of the metrics (heart rate, steps, sleep). Then we linearly interpolated the missing data for the remaining days. To leverage even more data, we kept the participants who had only 1 to 3 days missing data using a 24-hour template by averaging the activity for each hour across the time window (1 to 5 weeks).

D. Burnout Risk Prediction Models

We tested several models to predict burnout risk including Support Vector Machine (SVM), Generalized Linear Model (GLM), and Random Forest (RF) models. For the SVM, we used linear kernel, 'balanced' class weight, and set the regularization parameter of squared l2 penalty to be 0.1. The GLM we used is an Elastic Net regularized logistic regression model with a 'balanced' class weight. For the RF model, we used 100 estimators and set the class weight as 100:1 to give the minority class more weight.

We tested different combinations of features and window lengths to understand which type of features are more predictive of burnout, and how much time in advance we can make predictions. From the daily features, we made time windows that are organized into 7, 14, 21, 28, and 35 days of data, where a 7-days window refers to the first week of data, a 14-days window refers to the first 2 weeks of data, and so on. Depending on how the features were obtained, we classified them into 3 modalities: sensor (SN), survey (SV), and rhythm features with template imputation (RT). We aggregated the daily sensor and survey features across each time window. To select the most significant features, we used ANOVA (Analysis of Variance) F-value where our groups were high and low risk burnout participants. This allowed us to reduce the size of the rhythm features to the top 50 features in terms of their ability to disseminate between burnout risks. Min-max normalization was then applied to the dataset before prediction and analysis.

To fine-tune our models, we randomly divided our data into 5 folds and tested the model one-fold at a time, while the rest of the folds were combined for the training set. Then we used grid search to test our hyperparameters at each fold.

To evaluate the performance of our models, we used a leave-one-subject-out (LOSO) cross-validation approach. We tested the performance of a generalized model for each participant. We repeated this process 5 times to obtain average metrics for performance. We then used F1 score, sensitivity, specificity, and the area under the ROC curve (AUC) as our metrics.

E. Counterfactual Explanations

DiCE (Diverse Counterfactual Explanations) is a framework that generates diverse and feasible counterfactual explanations that go beyond traditional feature importance rankings or model approximations [20], [21]. DiCE offers insights into why specific predictions were made and can be used to explain any machine learning model in terms of feature perturbations. This promotes enhanced interpretability, transparency, and trustworthiness, making DiCE valuable for sensitive domains where explainability is crucial.

By employing counterfactual explanations through DiCE, we are able to gain a deeper understanding of the underlying reasons and effectively communicate them to the participants and healthcare providers. Using the optimal SVM model defined above, we applied our results to the DiCE framework, calculating 100 counterfactuals for each participant and assessing the alterations they exhibited compared to the original input. We first subtracted the original feature values from each of our counterfactuals (CFs) to demonstrate relative change. Using the CFs (n=3126) from our positive class (high-risk burnout) and their change to the negative class (low-risk

burnout), we clustered the data using kmeans clustering. By doing so we simplified the explanations into 6 clusters (k=6 was chosen based on silhouette score) and reported the 6 clusters centers to represent the average change in features.

III. Results & Discussion

A. Burnout Risk Prediction Model Evaluation

We compared the predictive performance of SVM, GLM, and RF models on the combination of all modalities of features (SN, SV, and selected RT) over the 35-day time window. All exhibited good predictive performance with F1 and AUC scores above 0.9. However, RF failed to correctly predict any negative class (low-risk burnout) sample despite of high AUC score, resulting in specificity to be 0. While SVM and GLM both displayed the ability to identify low-risk samples from the highly imbalanced dataset, SVM outperformed GLM in all metrics, achieving 0.99 in F1 score (0.97 with GLM), 0.97 in sensitivity (0.97 with GLM), 1.00 in specificity (0.80 with GLM), and 0.99 in AUC score (0.98 with GLM). Therefore, we chose SVM to be our model to evaluate the performance on different feature combinations and time windows, and the counterfactual explanations in the next section.

1) Compare different feature combinations: To understand which type of features are more predictive of burnout, we compared the performance of SVM on different combinations of modalities over the 5-week time window (Table I). The results demonstrate that rhythm features are important for the improvement of predictive performance, while feature selection is necessary to reduce the dimension and keep the models from learning spurious correlations of features.

TABLE I
PREDICTIVE PERFORMANCE OF SVM ON DIFFERENT FEATURES (5-WEEK TIME WINDOW) SN: SENSORS, SV: SURVEYS, RT: RHYTHM FEATURES

Features	F1	Sensitivity	Specificity	AUC
SN	0.67	0.53	0.60	0.60
SN+SV	0.81	0.72	0.60	0.67
SN+SV+RT	0.94	**1.00**	0.00	0.73
SN+SV+RT(selected)	**0.99**	0.97	**1.00**	**0.99**

2) Compare different time windows: To assess the extent to which we can make predictions in advance, we conducted a comparison of SVM performance across various time windows (Table II). Generally, predictive performance demonstrates improvement as the time window extends from 1 to 5 weeks, yielding the highest F1 score, specificity, and AUC score at the 5-week mark. Notably, the 3-week time window exhibited a considerable enhancement compared to the initial 1-week and 2-weeks and approaches the performance level of the 5-week time window. These findings suggest that effective predictions can be made up to 2 weeks in advance for our participants.

B. Counterfactual Explanation Evaluation

The resulting CFs clusters demonstrated 6 average CFs for changing burnout from high-risk to low-risk. Table III displays the 5 features from each cluster that change the most and the average scores of three burnout subscales for each cluster.

TABLE II
PREDICTIVE PERFORMANCE OF SVM ON DIFFERENT TIME WINDOWS; FROM $SN + SV + RT(selected)$

Week	F1	Sensitivity	Specificity	AUC
1	0.96	0.97	0.60	0.94
2	0.94	0.92	0.80	0.94
3	0.99	**1.00**	0.80	**0.99**
4	0.97	0.97	0.80	**0.99**
5	**0.99**	0.97	**1.00**	**0.99**

Total 21 unique features were found to be in the top 5 most changed values across all 6 cluster centers.

1) Analyze important features: The features 'sleep_20h_RSS (residual sum of squares)' and 'alcohol_mode35days' were the top 5 most changed features for 4 of the clusters. The feature 'sleep_20h_RSS' is a p-value that measures the goodness of fit test of the cosinor method on sleep data (see subsection 'Rhythm Features' above), and a low p-value implies good fitting of a cosine wave to the 20h-period data. For clusters 1, 2, 3, and 4, if the p-value drops the burnout risk level would change from high to low, which indicates that the participant should have a more regular 20 hour sleep rhythm in order to lower risk. The feature 'alcohol_mode35days' represents the most common number of cups of alcohol consumed each day over the 35 days. Our data show that for clusters 1, 3, 5, and 6, increasing this value moderately would lower the risk. This implies that moderate alcohol consumption would be helpful to reduce burnout risk.

The next most frequent features observed are 'sleep_diff_median23', 'hr_std_16h_phi' and 'sleep_diff_median27' which were found in 2 cluster centers each. 'sleep_diff_median23' and 'sleep_diff_median27' stand for the median of sleep time's hourly deviation from the template on the 23rd and 27th day respectively. As a decrease in these values would lower the burnout risk, it indicates that the more regular sleep patterns are, the lower the risk would be. The feature 'hr_std_16h_phi' is the acrophase of heart rate standard deviation fitted in the cosinor method on a 16h period, a measure of the time of overall high values recurring in each cycle. For clusters 4 and 5, the burnout risk level would reduce if 'hr_std_16h_phi' increases, which suggests that the later phase in the variability of heart rate (not HRV) in 0-24H, implying early sleep onset before midnight is negatively correlated with the burnout risk level.

Overall, 13 of all 30 features are related to sleep data and 11 are related to heart rate data, demonstrating the significance of sleep and heart rate features as markers for burnout risk prediction. Other than that, the reduction of overtime work and caffeine consumption demonstrates the importance of lowering burnout risk in clusters 1 and 4 respectively.

2) Associate clusters with burnout subscales: Cluster 1 exhibits comparatively low levels of emotional exhaustion and depersonalization, with a high personal accomplishment score. This suggests that CFs in cluster 1 generally come from participants who are likely positioned close to the decision boundary and have a relatively low risk of burnout. Cluster

TABLE III

TOP 5 FEATURE CHANGES FOR EACH CLUSTER: THE FEATURE NAMES THAT END WITH A NUMBER REPRESENT THE STATISTICS OF THE HOURLY DEVIATION FROM THE TEMPLATE ON THIS DAY. FOR EXAMPLE, HR_MEAN_DIFF_KURTOSIS1 STANDS FOR THE KURTOSIS OF HEART RATE MEAN'S DEVIATION FROM THE TEMPLATE ON THE FIRST DAY.

Cluster 1 (n=1384)		**Cluster 2 (n=1024)**		**Cluster 3 (n=262)**	
Features	***Values***	***Features***	***Values***	***Features***	***Values***
sleep_20h_RSS	-0.67	sleep_20h_RSS	-0.47	sleep_20h_RSS	-0.39
alcohol_mode35days	0.58	hr_mean_diff_kurtosis1	0.34	sleep_diff_median23	-0.15
sleep_onset_mean35days	-0.43	sleep_diff_median23	-0.32	hr_std_diff_median20	-0.13
sleep_diff_skew12	0.42	sleep_diff_kurtosis17	0.32	hr_std_diff_mean24	-0.13
overtime_work_mode35days	-0.36	hr_std_diff_median28	-0.28	alcohol_mode35days	0.12
Emotional Exhaustion	3.29 (0.33)	Emotional Exhaustion	3.83 (1.02)	Emotional Exhaustion	3.69 (0.67)
Depersonalization	2.69 (0.35)	Depersonalization	2.78 (0.98)	Depersonalization	2.57 (0.79)
Personal Accomplishment	3.48 (0.54)	Personal Accomplishment	2.11 (0.65)	Personal Accomplishment	2.30 (0.60)
Cluster 4 (n=198)		**Cluster 5 (n=144)**		**Cluster 6 (n=114)**	
Features	***Values***	***Features***	***Values***	***Features***	***Values***
hr_std_16h_phi	0.81	hr_std_16h_phi	0.70	naps_mode35days	0.75
sleep_diff_median27	-0.42	hr_mean_diff_kurtosis28	0.53	hr_mean_RA	0.64
sleep_20h_RSS	-0.36	hrmin_mean35days	-0.44	alcohol_mode35days	0.63
caffeine_cups_mode35days	-0.33	alcohol_mode35days	0.42	hrstd_mean35days	0.50
sleep_efficiency_mean35days	-0.27	sleep_diff_median27	-0.41	hrentropy_mean35days	-0.47
Emotional Exhaustion	3.72 (0.61)	Emotional Exhaustion	4.43 (0.37)	Emotional Exhaustion	4.18 (0.40)
Depersonalization	2.38 (0.55)	Depersonalization	2.84 (0.37)	Depersonalization	2.82 (0.52)
Personal Accomplishment	2.60 (0.65)	Personal Accomplishment	3.36 (0.98)	Personal Accomplishment	2.50 (0.19)

2 primarily relates to the personal accomplishment subscale, whereas clusters 5 and 6 are more closely associated with the emotional exhaustion and depersonalization subscales.

IV. CONCLUSION & LIMITATIONS

We predicted a high risk of burnout in shift workers using wearable sensor, survey, and rhythm features collected during a 5 week study. Our models identified rhythm features as crucial predictors, and we could accurately predict burnout up to two weeks in advance. Analyzing counterfactual explanations, we found that a more regular sleep rhythm could potentially reduce the risk of burnout syndrome. Our study also emphasized the importance of sleep and heart rate features.

There are several limitations and avenues for future work. Firstly, turning counterfactual insights into actionable plans typically requires the involvement of medical professionals. We plan to add more interpretable/modifiable features to the models. Secondly, our participants were mainly nurses, so further analysis should encompass shift workers in various roles for broader applicability. Additionally, we linked clusters to burnout subtypes, suggesting a need for a deeper exploration of variable correlations.

REFERENCES

[1] W.-J. Cheng *et al.*, "Night shift and rotating shift in association with sleep problems, burnout and minor mental disorder in male and female employees," *Occupational and Environmental Medicine*, vol. 74, p. 483, 7 2017.

[2] W. H. Organization, *QD85 Burnout in International Classification of Diseases, Eleventh Revision (ICD-11)*, 11th ed., 2019.

[3] M. Grzadzielewska, "Using machine learning in burnout prediction: A survey," *Child and Adolescent Social Work Journal*, vol. 38, pp. 175–180, 2021.

[4] K. Bauernhofer *et al.*, "Subtypes in clinical burnout patients enrolled in an employee rehabilitation program: differences in burnout profiles, depression, and recovery/resources-stress balance," *BMC Psychiatry*, vol. 18, p. 10, 12 2018.

[5] Y.-L. Lee *et al.*, "An app developed for detecting nurse burnouts using the convolutional neural networks in microsoft excel: Population-based questionnaire study," *JMIR Medical Informatics*, vol. 8, p. e16528, 2020.

[6] M. B. Hosseinabadi *et al.*, "The effects of amplitude and stability of circadian rhythm and occupational stress on burnout syndrome and job dissatisfaction among irregular shift working nurses," *Journal of Clinical Nursing*, vol. 28, pp. 1868–1878, 5 2019.

[7] W. H. Walker *et al.*, "Circadian rhythm disruption and mental health," *Translational Psychiatry*, vol. 10, p. 28, 12 2020.

[8] I. N. Karatsoreos, "Links between circadian rhythms and psychiatric disease," *Frontiers in Behavioral Neuroscience*, vol. 8, 5 2014.

[9] E. E. Kaczor *et al.*, "Objective measurement of physician stress in the emergency department using a wearable sensor," *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2020, pp. 3729–3738, 2020.

[10] H. Yu *et al.*, "Personalized wellbeing prediction using behavioral, physiological and weather data," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019.

[11] T. Iqbal *et al.*, "Stress monitoring using wearable sensors: A pilot study and stress-predict dataset," *Sensors*, vol. 22, no. 21, p. 8135, 2022.

[12] G. Vos *et al.*, "Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review," *International Journal of Medical Informatics*, p. 105026, 2023.

[13] M. T. Ribeiro *et al.*, "'why should i trust you?' explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[14] S. Wachter *et al.*, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[15] A. Ito-Masui *et al.*, "Internet-based individualized cognitive behavioral therapy for shift work sleep disorder empowered by well-being prediction: Protocol for a pilot study," *JMIR Res Protoc*, p. e24799, 2021.

[16] M. T, "The theory and measurement of burnout in human services," *The scientific report of Kyoto Prefectural University Humanistic Science*, vol. 11, pp. 99–112, 1987.

[17] J. R. Lunsford-Avery *et al.*, "Validation of the sleep regularity index in older adults and associations with cardiometabolic risk," *Scientific Reports*, vol. 8, p. 14158, 12 2018.

[18] M. Moškon, "Cosinorpy: a python package for cosinor-based rhythmometry," *BMC Bioinformatics*, vol. 21, p. 485, 12 2020.

[19] V. W.-S. Tseng *et al.*, "Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia," *Scientific Reports*, vol. 10, p. 15100, 9 2020.

[20] R. K. Mothilal *et al.*, "Explaining machine learning classifiers through diverse counterfactual examples," in *ACM Conference on Fairness, Accountability, and Transparency*, 1 2020.

[21] R. K. Mothilal *et al.*, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.